2015

Dysregulation of microRNAs in Blood as Biomarkers for Diagnosing Prostate Cancer

Rhonda W. Daniel
Virginia Commonwealth University

DYSREGULATION OF MICRORNAS IN BLOOD AS BIOMARKERS FOR DIAGNOSING PROSTATE CANCER

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Biochemistry at Virginia Commonwealth University

by

Rhonda Daniel

B.S. December 2012, Virginia Commonwealth University

Director: Zendra E. Zehner

Professor, Department of Biochemistry

Virginia Commonwealth University

Richmond, Virginia

July 2015

Dedicated to my amazing parents, Walid and Majida Daniel

Thank you for going against the social grain and encouraging me to spread my wings.

You two are truly amazing people. Everything that I am, I owe to you. I love you.

Acknowledgements

First and foremost, I want to thank my family for being such an amazing support system during this time and for putting up with the stress that comes with it. Without you all calling to check up on me, asking me how you can make things easier and just being in my corner, I would not have made it this far. Mom and dad, thank you for raising me to be independent and hardworking and allowing me to find my own way. You two have so much trust and faith in me and I hope you know just how thankful I am.

Second, I want to thank my committee. Dr. Zehner and Dr. Holmes, you both have been like my second mom and dad. The mentoring and advice I have received in the year and a half I have been with you are invaluable- you two are truly my biggest fans. Dr. Zehner, I cannot thank you enough for taking a chance on me. When Dr. Holmes volunteered me to work in your lab, nothing could have prepared me for the amazing work that we have done and for the priceless knowledge you have provided me with. And Dr. Holmes, THANK YOU for seeing something in me, and affording me the opportunity to work with your amazing wife! And thank you for sitting with me for hours and hours to discuss my future- you're always my loudest cheerleader! Dr. Sato-Bigbee, thank you for being so encouraging and excited about this work. Every time we meet it turns into a two hour affair and I appreciate it more than you know that you're willing to give me so much time to help me succeed. I am extremely grateful to all of you for you support, guidance, and inspiration.

I also want to thank Dr. Gail Christie for taking a chance on me and accepting me into the MBG program. I know that I came with my fair share of headaches but you have been a vital part of all of this and I cannot thank you enough for giving me this opportunity. I cannot count the amount of times I emailed you or barged into your office all panicked about one of the requirements or concerns that I had and you calmed me down. You have never been too busy to help or guide me and it means the world to me. They say when one door closes another one opens- you had this door wide open for me when all others were locked!

Last, but certainly not least, I want to thank my incredible fiancé. He has been a big support and has pushed me more than anyone else. Thank you for listening to me talk about things that really don't interest you, reminding me that there is life outside of work, and making me realize all the endless possibilities that lie ahead of me. I love you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ng	nanograms
µL	microliters
mL	milliliters
AGO	Argonaute
ALDH1A3	Aldehyde dehydrogenase 1 family, member A3
ARBS	Androgen receptor binding sites
AUC	Area Under the Curve
BCL6	B- Cell lymphoma
BPH	Benign prostatic hypertrophy
BTG	B cell translocation gene 1
CRPC	Castration- resistant prostate cancer
DHT	Dihydrotestosterone
DRE	Digital Rectal Exam
E2F	Eukaryotic transcription factors
FDR	False Detection Rate
GBM	Glioblastoma Multiforme
HDL	High density lipid
HK2	human glandular kallikrein
JAK	Janus kinase
KDM1A	Lysine- specific histone demethylase 1A

mRNA	messenger RNA
miR, miRNA	micro RNA
NPM1	Nucleolar phosphoprotein
pCa	Prostate Cancer
piRNA	Piwi- interacting RNA
PSA	Prostate Specific Antigen
qRT-PCR	Quantitative Real Time polymerase chain reaction
RISC	RNA induced silencing complex
RNA	Ribonucleic Acid
ROC	Receiver Operating Curve
siRNA	small interfering RNA
SKA2	Spindle and kinetochore associated complex subunit 2
STAT	Signal transducer and activator of transcription
TCGA	The Cancer Genome Atlas
TBF- beta	Transforming growth factor beta
TRPM	Transient receptor potential ion channels melastatin
VEGF	Vascular Endothelial Growth Factor

Abstract

DYSREGULATION OF MICRORNAS IN BLOOD AS BIOMARKERS FOR DIAGNOSING PROSTATE CANCER

By Rhonda Daniel, MS

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Biochemistry at Virginia Commonwealth University.

Virginia Commonwealth University, 2015

Major Director: Zendra E. Zehner, PhD. Professor, Department of Biochemistry & Molecular Biology

Prostate cancer is the most common noncutaneous cancer among men, yet current diagnostic methods are insufficient and more reliable diagnostic markers need to be developed. The answer that can bridge this gap and enable more efficient diagnoses may lie in microRNAs. These small, single stranded RNA molecules impact protein expression at the translational level and regulate important cellular pathways. Dysregulation of these small RNA molecules can have tumorigenic effects on cells and lead to many types of cancers.

Currently the Prostate-Stimulating Antigen (PSA) is used as a diagnostic marker for prostate cancer. However, many factors can elevate PSA levels such as infections and certain medications, consequently leading to false positive diagnoses and unnecessary

concern and over treatment with dire outcomes for the patient. Even worse, are the chances of false negative diagnoses, which result in prostate cancer not being diagnosed until its later stages. Therefore, although the use of the PSA level has had its uses in the clinic, it has failed to sufficiently bridge the gap or to distinguish indolent from aggressive disease.

It has long been suggested in the literature that microRNAs are drastically altered throughout the course of cancer progression. Here, RNA sequencing was used to identify changes in miR expression profiles diagnostic for prostate cancer patients compared to non-patient controls. The RNA sequencing results were also used to identify normalization miRs to be used as endogenous controls. Confirmatory qRT-PCR was then used to corroborate these results for the top seven dysregulated miRs found from the RNA sequencing data. Data analysis of the Area Under the Curve (AUC) of the Receiver Operating Curves (ROC) of the selected miRs exhibited a better correlation with prostate cancer (AUC Range= 0.819- 0.950) than PSA (AUC of PSA=0.667). In summary, a panel of seven miRs are proposed, many of which have prostate specific targets, which would represent a significant improvement over current testing methods.

Chapter ONE:

Introduction

I. The Prostate and Prostate Cancer (PCa)

Anatomy of the Prostate:

The prostate is an exocrine gland, and more accurately a muscle, associated with the male reproductive system. Located below the bladder and in front of the rectum, this gland is the size of a walnut[1]. The prostate surrounds the urethra, which allows it to secret an alkaline fluid that nourishes and protects sperm from the acidity of the vagina. This fluid comprises of about 30% of seminal fluid and contains many enzymes, such as the prostate- specific antigen (PSA), and hormone like substances such as spermine[2]. It also contains growth factors, lipids, zinc, potassium, citric acid, fructose, prostaglandins, etc. that enhance sperm survival in the harsh environment of the vagina[3].

This small gland is composed of three different cell types. The first type is the gland cell, which produces the alkaline fluid that is secreted during ejaculation. The second cell type is the muscle cell. During ejaculation, the sphincter muscles of the prostate gland, as well as the bladder, close the urethra to prevent backwash of semen into the bladder. The prostate also contains smooth muscle cells to help to expel semen during ejaculation. The final type of cells is the fibrous cell, which provides support for the prostate gland[4].

The prostate gland is also divided into three zones, a classification technique used mainly in pathology. The central zone is the innermost zone, which surrounds the ejaculatory duct. About 2.5% of prostate cancers arise from this area, but tend to be more aggressive and invasive. The transition zone surrounds the urethra and grows continuously

throughout the lifetime of men. Because this zone is constantly growing, benign prostatic hypertrophy (BPH) usually occurs in this zone and accounts for 10-20% of prostatic cancers. Finally, the peripheral zone, the outermost zone of the prostate, surrounds the distal urethra and can be felt by a physician during a digital rectal exam (DRE). This zone composes 70% of the prostate gland and about 70% of prostate adenocarcinomas arise from this zone[5].

The prostate can also be divided into several lobes- anterior, median, lateral, and posterior. The anterior lobe is the portion of the gland lying in front of the urethra. This part does not contain glandular tissue and is composed exclusively of fibromuscular tissue. For this reason, the anterior lobe rarely forms adenomas. The median lobe is situated between the ejaculatory ducts and the urethra. Contrary to the anterior lob, the median lobe consists of mainly glandular tissue and therefore is a common site of adenomas. On each side of the urethra lay the lateral lobes (right and left lobes), which form the main mass of the prostate gland. Since these lobes contain glandular tissue they are at risk of developing adenomas at an older age. Finally the posterior lobe is located closer to the rectum and can actually be palpated during a digital rectal exam (DRE)[6].

Diagnosis of Prostate Cancer:

The prostate specific antigen (PSA) is a prostate specific protein secreted by the epithelial cells of the prostate gland[7]. This antigen is actually a protease enzyme and is a member of the tissue kallikrein family, some of which are also prostate specific[8,9]. In fact, recently, another one of its family members, human glandular kallikrein (hK2) was discovered to be useful in detecting prostate cancer as well[9]. PSA is a major component of semen where its main function is to cleave semenogelins in the seminal coagulum, which liquefies semen to allow sperm to swim freely[8]. It is instrumental in dissolving the cervical mucus and allowing for the entry of sperm into the uterus.

PSA and other kallikreins can be tissue or plasma kallikreins, which cleave peptide bonds. There are 15 known tissue kallikreins all located in the same gene cluster on the long arm of chromosome 19. Interestingly, a study conducted in 2003 showed that there were sections on chromosome 19 that contained regions of tumor aggressiveness in prostate cancer[10,11]. This suggests that the alleles present in this area dictate the aggressiveness of disease, specifically prostate cancer.

About 30% of PSA in seminal plasma is the intact proteolytically active enzyme and 5% is complexed with a protein C inhibitor. However, the majority (70%- 90%) of PSA enter the peripheral blood complexed with a protease inhibitor, which catalytically inactivates it. PSA can also circulate in the blood in its free form (10%-30%)[8]. The ratio of free to total PSA (PSA index) has been found to be lower in patients with pCa, which aids in differentiating between normal and pCa[8]. To date total PSA levels of less than 4 ng/uL in blood are considered normal while higher levels have been associated with

prostate cancer[12]. However, the literature has shown that this is not always true. In one study, 18,882 men had their PSA levels measured and of those 2,959 were found to have PSA levels that are considered within the normal range. Among those men, prostate cancer was diagnosed in 449 individuals, 67 of which had a Gleason score of 7 or higher. This study showed how unreliable the use of PSA level as a diagnostic tool was in prostate cancer[13]. Furthermore, even high PSA levels may not indicate prostate cancer. In fact, studies have shown that 75% of men with elevated PSA levels do not have prostate cancer[14].

Since this method of diagnosis is clearly unreliable, the only way to make a definitive diagnosis is by a prostate biopsy. This requires a sample of body tissue to be removed from the prostate and examined under a microscope by a pathologist to determine if abnormal cells are present. Only then can a conclusive diagnosis be made[15].

Grading Prostate Cancer:

After diagnosis of prostate cancer, a grading score, called a Gleason score, is given based on the morphology of the cells seen by the pathologist. Gleason scores range from 2 to 10 and suggest the prognosis of the prostate cancer and how likely that tumor will spread. The lower the Gleason score the more similar the cancer tissue is to the normal prostate cells. As the score increases, the cancer cells become more distinguished from their normal counterparts and increase the likelihood of the tumor metastasizing. A score of 2 to 4 indicates that the cancer cells share similar characteristics to the normal cells of the prostate and are associated with a good prognosis. A score of 5 to 7 denotes intermediate risk with cancer tissue becoming more distinguished from normal tissue. Finally, a score of 8 to 10 suggests little to no similarities between the cancer and normal tissue and tend to be more aggressive in their nature[16].

A detailed look at the prostate biopsy can break this score down further into two parts. For example, a Gleason score of 8 with a break down of 3+5 indicates that the most common tumor grade pattern seen by the pathologist was 3 while the second most common was 5. This individual would have a better prognosis than an individual with a Gleason score of 8 but with a break down of 5+3 because the most abundant grade pattern was a 5 indicating that the majority of abnormal cells were morphologically more different than the normal cells[16].

II. MicroRNA Biogenesis and Role in Disease

MicroRNA Biogenesis:

Small RNAs are an extremely important part of gene regulation. Their role in suppression of unwanted genetic materials and transcripts is monumental to the proper function of the body. Small RNAs are small in length (20-30 nts) and are associated with Argonaute family proteins (AGO family proteins). They fall into three classifications microRNAs, siRNA, and PIWI- interacting RNA (piRNA), the most dominating of which are microRNAs[17]. MicroRNAs (miRs) are small non-coding RNA molecules, between 18-22 nucleotides, that have been evolutionarily conserved. These miRs function at the translational level through silencing mechanisms to regulate gene expression. The first miR was discovered in the 1990s but it was not until the 2000s that they became recognized as a class of biological regulators[18].

More than half of known miRs are located within introns of protein coding or noncoding transcription units. However, 10% are encoded by exons of long nonprotein-coding transcripts. Biogenesis of microRNAs begins in the nucleus and continues in the cytoplasm of cells. Transcription is carried out by RNA polymerase II, the same enzyme that transcribes messenger RNA (mRNA) from DNA, to produce a pri-miRNA transcript that is about 70 base pairs long. The pre- miRNA has a characteristic stem and loop structure and can encode for more than one miR precursor. Drosha, a class II ribonuclease III enzyme located in the nucleus, cleaves this precursor further so that it can be transported by Exportin-5 to the cytoplasm. Pasha, an RNA recognition protein thought to

aid in template recognition, assists Drosha. A second ribonuclease III enzyme, DICER, cleaves the precursor once more to produce the mature form of the microRNA with the help of another RNA recognition protein, TRBP. The mature miR then associates with the RNA- induced silencing complex, RISC, which allows it to carry out its gene silencing functions[18].

The first 2-8 nucleotides of the miRNA, termed the seed region, determine which mRNA is targeted[19]. The seed region must be complementary to the 3' untranslated region (3' UTR) of the mRNA in order for silencing to occur. However, there have been examples of miRs binding to the 5' UTR region of the mRNA. It is also worth noting that there are examples of miR targets where the seed region is not completely complementary. Similarly, a complementary sequence between the microRNA and mRNA does not always mean that it is a target. Many mRNAs are predicted to be miR targets, but validation is required experimentally before such a conclusion can be reached. Therefore, miRNAs can have as many as 100 predicted targets[20].

MiR gene silencing mechanisms depend on how complementary the miR sequence is to its target mRNA. If the miR sequence is completely complementary to its mRNA target, degradation occurs. However, if the sequences do not completely match, suppression of mRNA translation ensues by acting as a block that prevents the ribosome from translating the mRNA into a protein[20].

Circulating microRNAs:

MicroRNAs have been found to be secreted from the cell for signaling in a paracrine, autocrine or endocrine manner. These miRs have been found to be extremely stable and have excited researchers as possible diagnostic biomarkers for disease. Circulating miRs can be stored in microvesicles, exosomes, in protein complexes bound to Ago-2 or NPM1, or in HDL particles bound to Apo[21-26].

The purpose of secreted microRNAs depends on their mode of transportation. For example, miRs secreted in vesicles appear to be secreted as a means of communication since they do not represent the miR profile within the cells[26,27]. Interestingly, precursors of many miRs are also found in these compartments[28].

Since their discovery many research groups have analyzed the microRNA composition of blood for various disease studies. Schultz et al, studied whole blood for the identification of microRNAs that could be used as biomarkers for the detection of pancreatic cancer. By confirmatory qRT-PCR, they found 38 microRNAs dysregulated in whole blood including hsa-miR-150, hsa-miR-636, hsa-miR-145, hsa-miR-223, hsa-miR-26b, hsa-miR-34a, hsa-miR-122, hsa-miR-126, and hsa-miR-885. They were able to identify 2 diagnostic microRNA panels that could distinguish to a certain degree between patients with pancreatic cancer from healthy controls[29].

Another study compared microRNA levels between plasma and serum by measuring 4 microRNAs: hsa-miR-15b, hsa-miR-16, hsa-miR-19b, and hsa-miR-24. Interestingly, they found that there was a strong correlation in the microRNA content of the two types of [30]. Therefore both of these means would be sufficient in disease study.

A study conducted by Andreas Keller et al, used RNA sequencing of whole blood samples collected with PAXgene Blood tubes to study microRNA profiles in lung cancer patients. However, samples were pooled together prior to sequencing, therefore, preventing individual sample analysis. As a result none of these studies used RNA sequencing to analyze miR dysregulation across different individual samples.

MicroRNAs in Disease:

Aberrant expression of microRNAs has been associated with various diseases. Since they play an important role in gene silencing, over expression or under expression can lead to dysregulation of certain pathways. This is especially true in cancer where abnormal cell growth and angiogenesis are critical for the tumorigenesis to occur. The loss of microRNAs that suppress the translation of oncogenes, termed tumor suppressors, has been shown in many cancers. These miRs are primarily responsible for apoptotic pathways and cell cycle checkpoints[31]. Our lab previously identified miR-125b as a tumor suppressor lost in the M12 prostate cell line which is highly tumorigenic and metastatic, compared to the poorly tumorigenic and not metastatic parental P69 prostate cell line. By restoring miR-125b, the migratory and invasive capability of the M12 subline was significantly reduced[32]. Conversely, some miRs target tumor suppressor genes and may become upregulated in cancer. These miRs, known as oncomiRs, allow for abnormal cell growth and angiogenesis by inhibiting apoptotic pathways and hindering cell cycle checkpoints. Similarly, in the M12 subline we have documented an up regulation of miR-22, which targets the cellular brake PTEN, thereby contributing to enhanced tumor growth[32,33]. miR-27a has also been labeled an oncomiR in the literature by targeting prohibitin which results in a decrease in the translation of its mRNA and promotes an increase in prostate cancer cell growth[34].

III. MicroRNAs as Diagnostic Markers for Prostate Cancer (Project Aims)

Aim 1. To identify diagnostic microRNAs in prostate cancer patients using RNA sequencing. Hypothesis: Dysregulation of microRNAs leads to cancer. Therefore, microRNAs will be upregulated or downregulated throughout the progression of prostate cancer. RNA sequencing of microRNA extracted from whole blood will allow for the identification of all miRs present in the miR expression profile of patients and controls as well as provide insight as to which miRs may play a critical role in the progression of this disease. Furthermore, it will provide an idea of how miRs are acting (whether they are being upregulated or downregulated) as the disease progresses.

Aim 2. To identify a panel of miRs in blood to serve as normalizers for qRT-PCR. Hypothesis: MiRs with constant expression across all the samples can serve as endogenous controls, which can help to identify dysregulation of miRs that could be diagnostic for prostate cancer.

Aim 3. Use confirmatory qRT- PCR to substantiate RNA sequencing results as well as determine changes in microRNA expression profiles as prostate cancer progresses.

Hypothesis: with all the advancements being made in the science field, qRT-PCR remains the most effective and sensitive method in determining miR presence. Thus, qRT- PCR of the same RNA samples used for RNA sequencing will be used to confirm miR dysregulation. For the most part, miRs chosen for validation will be based on those shown to be dysregulated via RNA sequencing analysis and supported by the literature where possible.

Chapter TWO:

Methods and Materials

This chapter is a comprehensive discussion about the techniques used to isolate and analyze RNA from samples and controls. A total of 12 controls and 28 patient samples were extracted and analyzed using the methods discussed below. Various limitations and advantages for each technique are discussed as well.

I. Extraction

Protocol:

Plasma and whole blood samples from patients and normal individuals were obtained from the Nelson Urology Clinic and the Hunter Holmes McGuire VA Medical Center in compliance with their respective IRBs. Plasma samples were collected in purple- capped EDTA tubes to prevent cell lysis and whole blood samples were collected in PAXgene blood collection tubes (PreAnalytiX, a Qiagen/ BD company). PAXgene tubes contain a reagent that lyses blood cells and immediately stabilizes intracellular RNA to prevent degradation. For this reason, the tubes were left at room temperature for a minimum of two hours after collection. Plasma was spun down at 3,000 rpm as soon as the sample was collected to prevent cell lysis and the supernatant was then placed in 1.5 mL microtubes in 500 μL aliquots to be stored at -80C.

A corresponding extraction kit, PAXgene Blood microRNA kit, was used to extract microRNA from the PAXgene collection tubes according to the manufacturer's instructions (PreAnalytiX, a Qiagen/ BD company). This kit allowed for preferential extraction of RNA within the length parameters of microRNAs using a silica membrane.

The isolated RNA was eluted in 80 µL of elution buffer and stored at -80C.

Supernatant aliquots were processed using the miReasy Plasma/ Serum extraction kit according to manufacturer's instructions (Qiagen). Aliquots were complied onto a single RNeasy® MinElute® Spin Column and eluted in 25 µL of elution buffer. All RNA samples were stored at -80C.

Principle:

RNA is susceptible to rapid degradation by nucleases present in blood within hours of blood extraction. Furthermore, gene regulation, such as induction and repression, can lead to changes in transcript levels, which can change the overall RNA profile. PAXgene collection tubes combat this issue by the inclusion of a liquid reagent that stabilizes intracellular RNA to prevent changes in the gene expression profile[24].

The miRNEasy Serum/ Plasma kit was designed for the purification of small RNA such as miRNA, from small volumes of serum. It uses a silica- membrane- based purification method to isolate and purify total RNA. The QIAzol Lysis thiocynate facilitates the lysis of cells, denatures proteins and RNases. Addition of chloroform allows nucleic acid, both RNA and DNA, to be efficiently partitioning in the aqueous solution and denatured proteins to be discarded in the interphase. The addition of ethanol at 0.5X volume of aqueous volume precipitates RNA, which is then preferentially bound to the spin column so that other contaminants can be washed away.

Limitations:

Several problems arose during the course of experimentation. Initial PAXgene blood tubes collect whole blood in a stabilizing solution, but the subsequent miR extraction kit appears to enhance for microRNAs. Although this increase in the yield of miRs results in sufficient RNA for sequencing analysis, this technique would not be ideal to use for studies interested only in analyzing circulating microRNAs. However, we could not obtain sufficient RNA for next generation sequencing from the usual small amounts of patient plasma necessitating starting with whole blood. Nonetheless, theoretically, any cellular miR contamination that is seen in the patient profiles will be comparable to the patient profile IF it is not due to disease. As a result, it was felt that for the purposes of this study, whole blood would suffice because regardless of the source of the miR (cellular or plasma), if it turns out to be diagnostic for prostate cancer, it's useful. Furthermore, at this time it is not clear which miR particle- exosomes, microvesicles, apoptotic bodies, or bound to proteins such as Arg-, or HDL- would be most diagnostic for prostate cancer detection. Thus at this stage, our approach is not limited to any one particle, which could jeopardize results, rather we are including all particles in our analysis. Future experiments, will be needed to determine if any one of these different types of microRNA containing particles is superior to the other in predicting disease.

Nonetheless, to begin an investigation into the possibility, a few plasma samples were also collected in order to compare the circulating miR profile of patients and controls with their corresponding whole blood profile. This allowed for a comprehensive look at what miRs are being excreted and what miRs are due to cellular contamination.

Furthermore, enhancement for some miRs could potentially skew the miR profile in favor of overly expressed miRs while wiping out the documentation of less abundant miRs. This possibility will be further discussed in the next chapter, Chapter 3, as this is also a concern for bias in analyzing RNA sequencing results.

II. Bioanalyzer:

Protocol:

Samples were stored in 2 µL aliquots and heat denatured for 2 minutes at 70 °C. Bioanalyzer analysis was run according to manufacturers protocol using the pico chip and smallRNA chip assay (Agilent).

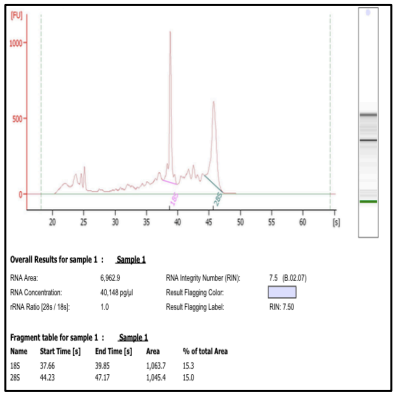
Principle:

The bioanalyzer allows for the quantification and qualification of extracted microRNA. The small RNA chip contains 11 sample wells, 3 gel wells, a conditioning solution well, and an external standard (ladder). The cells are interconnected to form an integrated electrical circuit. The bioanalyzer has 16 pin electrodes that fit in the 16 well chip, each with its own power supply, which allows for each sample to be run independently. A voltage gradient is created in each well causing charged molecules such as RNA to migrate, much like gel electrophoresis. The constant mass-to-charge ratio and sieving polymer matrix allows the molecules to be separated based on their respective size. The ladder includes six fragments ranging from 25 nucleotides to 150 nucleotides at 150 ng/ µL. By comparison to the ladder, miR quantification and size can be determined. When the fluorescent dye intercalates within the RNA the detector senses the RNA and determines the size and quality.

The advantage of running the pico chip lays in the fact that it shows small RNA as well as ribosomal RNA. Identification of the ribosomal contamination helps determine

how much of the sample is actually needed for RNA sequencing which requires 100 ng/ μL of total RNA **(Figure 2-1A)**. However, the small RNA chip is able to separate small RNA more effectively to better determine a microRNA concentration within the 10 nt- 40 nt region **(Figure 2-1B)**. Interestingly, we decided to determine the comparability and reproducibility between the two chips and found that the concentrations were remarkably close. To calculate the small RNA concentration in a sample run on the pico chip, the ribosomal contamination was subtracted from the total RNA concentration. The resulting number was amazingly close to the small RNA concentration determined by the small RNA chip.

A.



B.

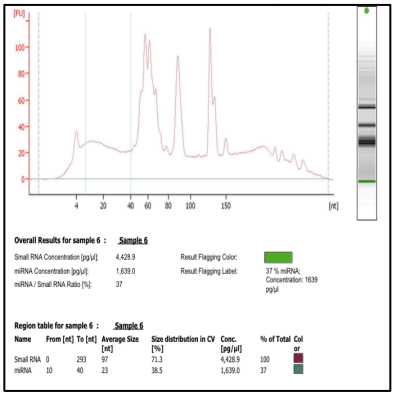


Figure 2-1 Representative Patient Comparison of Bioanalyzer Results from Pico Chip and Small RNA Chip: *Whole blood (2 uL) was analyzed using the pico chip and the small RNA chip. This is a representative patient bioanalyzer comparing between the two chips.* ***(A):*** *Pico Chip:* The 18S and 28S show the ribosomal concentration present in the sample. The percentage of contamination is reported here as 30.3%. Therefore to determine the small RNA concentration the concentration of ribosomal contamination was subtracted from total RNA concentration (4853.9 pg/ uL). This number is remarkably close to the small RNA concentration that is seen in the small RNA chip. **(B):** *Small RNA Chip:* Small RNA concentration as well as microRNA concentration are reported here.

Limitations:

The bioanalyzer, while useful at determining the approximate concentration of microRNAs, does not allow for exact quantification. It analyzes a range between 10-40 nt as the microRNA range although it is known that miRs fall within the 18-20 nt range. As a result the concentration reflected in the bioanalyzer results is an overestimate of the microRNA concentration in a given sample. Furthermore, the small RNA chip does not allow for the determination of ribosomal contamination. Although the pico chip was originally used and the ribosomal contamination was determined to be minimal, it was impossible to estimate ribosomal contamination for the samples seen on the small RNA chip alone. Prohibitive costs did not allow all samples to be analyzed on both chips and since a determination of miR concentration was the more important value to determine, all samples were analyzed on the small RNA chip.

III. RNA Sequencing:

Protocol:

100 ng/ μL of each sample was sent to the Core for sequencing using the Illumina TruSeq Small RNA Library Prep platform according to manufacturers protocol.

Principle:

RNA 3' adapters are ligated onto each sample followed by the 5' adapter to bring the small RNA size to ≅ 70 base pairs. For pair end sequencing, two indices are added to the sample as well to allow for the forward and reverse sequencing of the template strand. The RNA sample is then run on a gel to separate RNA based on size. At this point, the desired size (microRNA with ligated adaptor sequences and index sequences) falls in the 145 base pair region. Reverse transcription polymerase chain reaction (RT-PCR) is the carried out to synthesize a single stranded cDNA. The cDNA is then PCR amplified using a common primer and a primer containing 1 of 24 index sequences, which allows for the separation of the indexes and the RNA ligation reaction. "This design allows for the indexes to be read using a second read and significantly reduces bias compared to designs that include the index within the first read." The resulting cDNA is then gel purified and extracted. Once these isolated cDNAs are applied to the flow cell, there is cluster amplification.

Limitations:

Although RNA sequencing is a useful method because it allows scientists to determine what miRs are being expressed and the pattern of expression, it remains limited in some aspects. Analysis between batches is difficult due to "batch effects." These effects are inherent in any batch due to technical sources of variation that can be added to the samples during handling. Therefore comparison between two batches requires statistical analysis to correct for these differences. Furthermore, the various techniques that can be used in RNA sequencing such as paired end analysis or shotgun (single end) sequencing prevents the pooling of two batches for dual analysis. Therefore, it is important to be mindful of these limitations during the analysis of different RNA sequencing batches in order to prevent the addition of biases.

IV. qRT- PCR

Protocol:

Three ng of each RNA sample was converted into 20 µL cDNA using the qScript™ cDNA synthesis kit according to manufacturers protocol (Quanta). The cDNA was then diluted 1:2 and 2 µL were used for each PCR reaction run in triplicate. PCR was conducted on several miRs (Quanta Primers) discussed in the subsequent chapters using the qScript™ One- Step SYBR Green qRT-PCR kit according to manufacturers protocol (Quanta).

Principle:

The SYBR Green mix used contains all the components needed for the polymerase reaction including dNTPs, magnesium chloride, AccuStart Taq DNA polymerase, stabilizers, and SYBR Green I dye. When added to the cDNA sample and heated, denaturation yields a single strand as well as liberation of the DNA polymerase from its stabilizers (95 °C for 2 minutes). Upon cooling to the optimal annealing temperature (60 °C for 30 seconds), the primer can then anneal to the template strand and elongation (extension) can take place (70 °C for 30 seconds). This cycle is repeated for 40 cycles. At the end of each extension cycle the amount of SYBR Green bound to the double strand DNA product emits a fluorescent signal. The amount of this signal is quantitated, and the signal intensity increases each cycle as the PCR product accumulates. Once this signal reaches a preset threshold value the cycle number is noted. This threshold value is set to

be within the optimal amplification range of the cDNA sample. The lower the Cycle threshold (CT) value the more abundant the primer, and hence the miR, is in the samples. Vice versa, the higher the CT value, the less expressed the miR is in the sample.

Chapter 3

Next Generation Sequencing (NGS) Analysis

Introduction:

Next Generation Sequencing (NGS) is a relatively new method for the development of miR profiles. High- throughput NGS platforms provide researchers with the ability to analyze whole miR expression profiles with low amounts of microRNA input and quick output of results. This is a far cry from the First GS methods such as Sanger sequencing, which require cloning of one miR, amplification of that single miR and finally sequencing one base at a time. In contrast, the Illumina high- throughput platform and its sister platforms allow millions of reads to be sequenced at the same time.

A major advantage of RNA sequencing is its ability to distinguish miRs with even single nucleotide differences, such as the Let family. Since these differences fall outside the seed region of the miRNA, the targets of members of this family can differ vastly and the clinical implications of dysregulation change[33]. As a result, the fact that NGS can distinguish between these family members allows researchers to determine which miRs are important in disease. Furthermore, NGS can discriminate between isomiRs, which are a result of post-transcriptional additions to the 3' end, and less frequently the 5' end, of mature miRs. Another major advantage of NGS is identification of the whole miR profile of a sample. All miRs that are present in the sample will be represented in the sequencing profile. As a result, researchers can determine which miRs are worth validating with qRT-PCR.

RNA sequencing does, however, have its disadvantages. Despite being much more expensive than other methods, it still does not provide the sensitivity of qRT-PCR. For this reason validation of the miR expression profile is required. Moreover, since there are a limited number of oligonucleotides to bind to the adaptor sequences within a sample lane, microRNAs

that are abundant may be biasedly expressed in the sequencing profile than less abundant miRs. Therefore, when analyzing the sequencing results, it is important to keep these limitations in mind or the project runs the risk of excluding important miRs. Finally, significant adaptor trimming, genome alignment, and statistical analysis are required after resulting sequencing reads are obtained.

As discussed in the Methods and Materials section, the Illumina platform was used to analyze the miR profiles of controls and patients. Two different sequencing methods were available, including shotgun sequencing and pair-end sequencing. Shotgun sequencing is a quick method that allows for rapid output of results. Large fragments are sheered into smaller fragments about 30 to 350 nt long, which make it easier to sequence. The fragments are then reassembled into their original order based on overlapping sequences, which ultimately yields the complete sequence. However, this method runs the risk of producing unknown sequences at certain positions and is not very reliable when analyzing longer sequences. For the purposes of microRNA studies, literature suggests that while this is much less of a problem due to the small length of microRNAs, small sequencing errors were still seen in the results[35].

Pair- end sequencing, however, is more rigorous due to the sequencing of both ends of the fragment as discussed in the Methods and Materials section. Therefore, high- quality reads are generated which increases the likelihood of proper alignment to the reference genome.

Analysis:

For this study, three separate RNA sequencing batches were sent to the Nucleic Acids Research Facility at Virginia Commonwealth University. In total, 12 control samples and 28 patient samples **(Table 3-1)** ranging from elevated PSA to G9, with one sample from a patient with metastatic disease (*as noted) were studied **(Table 3-2)**. When age, race/ethnicity, or PSA level was provided, they are reported here. Furthermore, although the majority of the patient samples collected were G7-G8, an almost equal number of lower grade prostate cancer with good prognosis (G2-G6) and higher grade prostate cancer with poor prognosis (G7-G9) were obtained with 16 samples and 12 samples in each group respectively. Paired end sequencing using the Illumina platform was conducted as stated in the methods and materials section. Analysis of the RNA sequencing results was then preformed using multiple programs including Partek, RStudio, and edgeR.

Partek® Flow® is an online software program developed by Partek incorporated that analyzes RNA sequencing data. A pre-alignment QA/QC was run on the unaligned reads prior to analysis to determine the quality of the NGS raw data **(Figures 3- 1A and 3- 1B)**. The Average base quality score per position **(Figure 3-1A)** shows that the quality of bases at positions 16 to 45 have the highest quality scores. The Average base quality score per read **(Figure 3- 1B)** indicates that the Phred quality score is not uniform among the reads. Therefore, additional trimming of the adaptors and bases was necessary to increase the quality of the raw data.

Flow® was then used to trim the adaptor sequences (AGATCGGAAGAGCACACGTCT) ligated to the 3' and 5' ends of the templates as used by the Illumina sequencing TruSeq Small RNA sample preparation kits. The default settings were used except for determining the parameters for *match times* and the *min read length*. In some instances the adaptor sequence may anneal more than once so changing the *match times* from 1 (default) to 3 allowed the sequence to be cleaved up to 3 times from a single read. The *min read length* was changed from 25 to 16 since microRNAs are 18-22 nucleotides long.

Table 3- 1: Sample Log List

Sample Name	Control or Patient	Gleason Score	PSA Level	Age	Race/ Ethnicity	Bioanalyzer (ng/ µ L)
03132014a	PATIENT	7	-	-	-	4.77405
03132014b	PATIENT	7	-	-	-	4.2512
03132014c	PATIENT	7	-	-	-	6.25885
03132014d	PATIENT	9	-	-	-	2.34317
04292014!	PATIENT	-	Elevated	-	-	2.6643
05132014!	PATIENT	-	Elevated	-	-	4.81057
05242014!	PATIENT	9	-	92	Caucasian	7.48465
06062014!	PATIENT	-	22	-	-	2.41735
09182014!	CONTROL	-	-	-	Caucasian	1.01513
10212014	CONTROL	-	-	-	-	1.57365
12172014a	PATIENT	7	-	-	-	3.531
12172014b	PATIENT	-	Elevated	-	-	6.3641
12192014	CONTROL	-	-	51	Caucasian	4.74255
02162015!	CONTROL	-	-	24	Caucasian	11.1381
02162015!	CONTROL	-	-	24	Caucasian	5.7136
GG	CONTROL	-	-		Caucasian	13.07285
40 P.Y.J	PATIENT	7	3.42	66	Caucasian	2.49433
36 K.A.B	PATIENT	7	10.73	55	Caucasian	3.79373
18 E.H.D	PATIENT	6	6.8	68	African American	278.41495
9 Q.M.A	PATIENT	6	8.82	55	Asian/ Hawaiian	0.82855
72 W.J.C	PATIENT	8	4.77	63	Caucasian	2.6045
100 C.T.L	PATIENT	7	7.78	61	African American	1.5708
56 J.D.J	PATIENT	7	4.57	68	Caucasian	5.0533
85 D.T	PATIENT	7	6.1	65	Caucasian	1.9734
33 H.E.W	PATIENT	6	3.2	63	Caucasian	1.7158
20 A.O.E	PATIENT	6	4.15	64	African American	0.1449
31 E.M.R	PATIENT	6	12.47	67	Caucasian	0.183
62 M.A.V	CONTROL	-	-	39	African American	6.2787
51 D.L.D	PATIENT	7	6.06	64	Caucasian	13.801
6 W.H.C	PATIENT	6	11	58	African American	5.8854
83 R.L.J	PATIENT	6	4.21	66	Caucasian	12.8379
16 E.W.J	PATIENT	8	7.99	76	Caucasian	2.5383
82 J.R.D	PATIENT	7	8.36	62	African American	7.79565
8 R.M.S	PATIENT	6	4.4	73	African American	10.2831
89 B.R.W	PATIENT	7	3.5	66	Caucasian	10.7295
03262015 K.B	CONTROL	-	-	43	Caucasian	2.1404
04062015 W.L	CONTROL	-	-	91	Caucasian	15.8662
04242015 M.A	CONTROL	-	-	59	Caucasian	2.0354
04282014 S.S	CONTROL	-	-	55	Asian/ Hawaiian	2.453

Table 3- 2: Number of Samples for each Prostate Grade

Prostate Cancer Grade	Number of Samples
Elevated PSA	4
G3- G4	0
G5- G6	8
G7-G8	14
G9 + Mets	2

A. **B.**

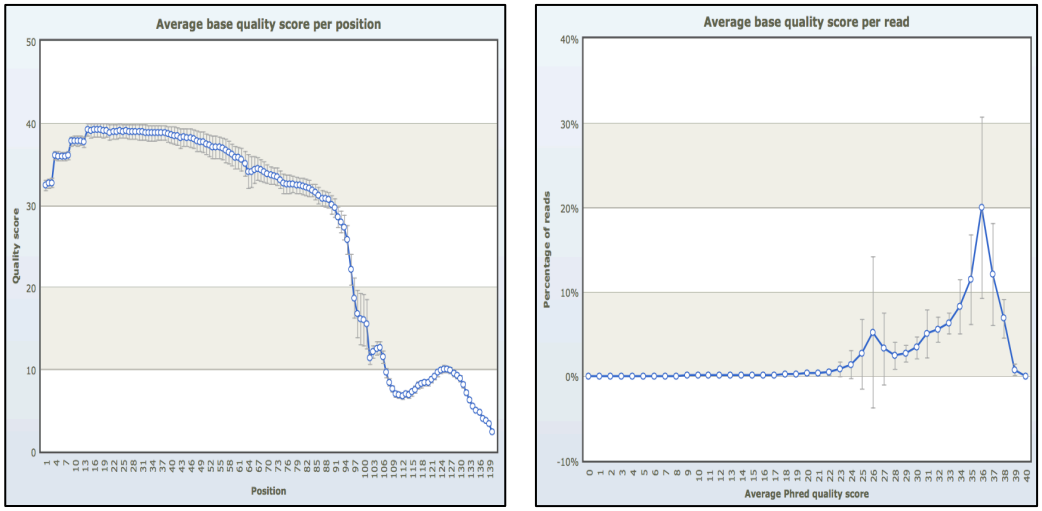


Figure 3- 1: Pre- alignment Analysis of QA/QC- Raw Data: (A): *Pre- alignment QA/QC of Average base quality score per position.* This graph shows the average quality score of each base. It is obvious that the bases at the beginning (Positions 1-16) and the bases at the end (46-139) have very low quality scores indicating a need to trim these bases off. **(B):** *Pre- alignment QA/QC Average base quality score per read.* This graph shows the quality score per read. This indicates that about 5% of the reads have a phred score of 26 suggesting that trimming is required to ensure that all the reads have quality scores of 38 or greater.

Another pre-alignment QA/QC report was generated to determine if the adaptor sequences were properly cleaved **(Figures 3- 2A and 3- 2B)**. The two graphs indicate an increase in the quality of the bases and the overall reads but further trimming was required. Trimming of bases on both ends of all the reads permits removal of partial sequences that may have been inserted into the template during cDNA library prep and amplification. Removal of these "junk" sequences allows for increased quality (Phred quality scores) of the reads and trims away bad quality bases that could affect alignment. A final pre-alignment QA/QC report was generated **(Figures 3- 3A and 3- 3B)** to show that all the bases had similar quality scores and that the average base quality for the majority of the reads was now at a Phred score of 39, where a Phred score of 30 is considered exceptional. This score indicates that there is a 1 in every 7,943 chance of an incorrect base call and that the base call accuracy is 99.99%. This pre-alignment data was convincing that the data was of high quality and that the analysis could continue.

Next the raw data had to be aligned to the human genome to determine the chromosome number as well as the start and stop positions of each read compared to the reference genome. The Bowtie aligner was utilized using the updated Genome Reference Consortium (GRCh38), with three parameters altered. *Seed mismatch limit* was altered from the flexible 2-mismatch value to a more stringent value of 1-mismatch, which ensured that any read that aligned to the genome was in fact a match.

A. **B.**

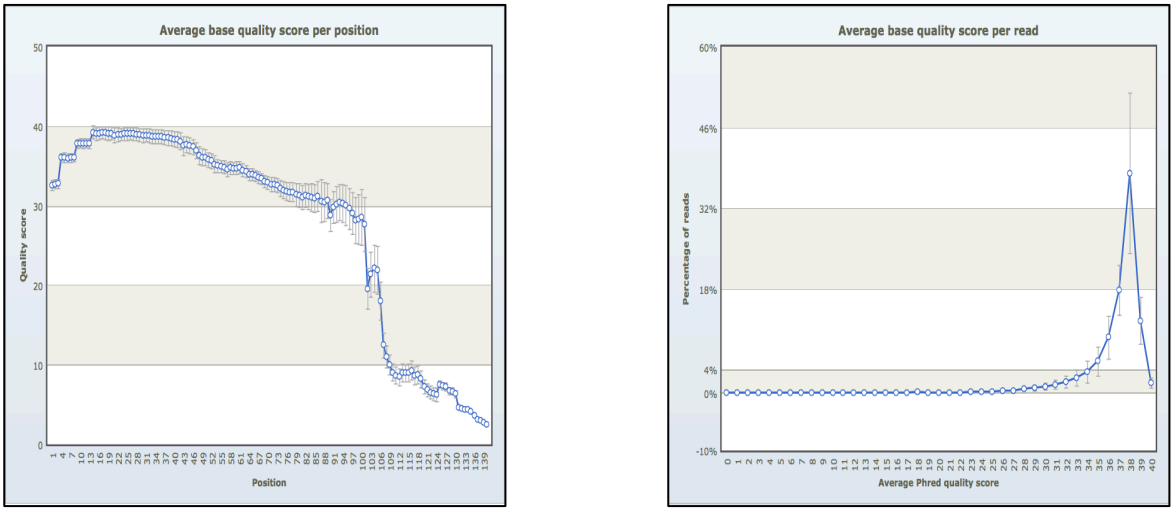
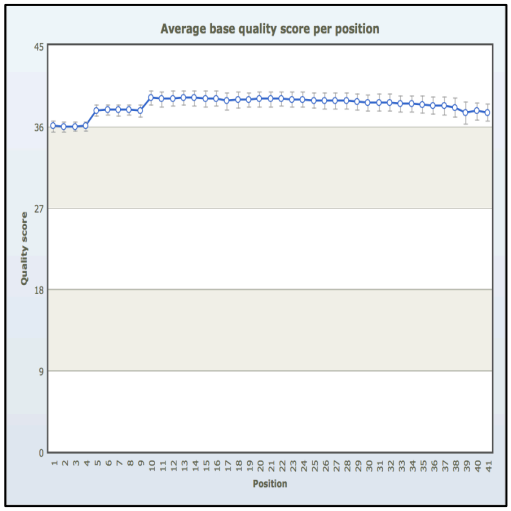


Figure 3- 2: Pre- alignment QA/QC- Adaptors Trimmed: (A): *Pre- alignment QA/QC of Average base quality score per position after Trimming of Adaptor Sequences.* This graph shows the average quality score of each base. It is obvious that the bases at the beginning (Positions 1-13) and the bases at the end (46-139) have very low quality scores indicating a need to trim these bases off. **(B):** *Pre- alignment QA/QC Average base quality score per read after Trimming of Adaptor Sequences.* This graph shows the quality score per read. This indicates that more than 100% of the reads have a Phred score higher than 31 with the majority of reads having a Phred score of 38.

A.



B.

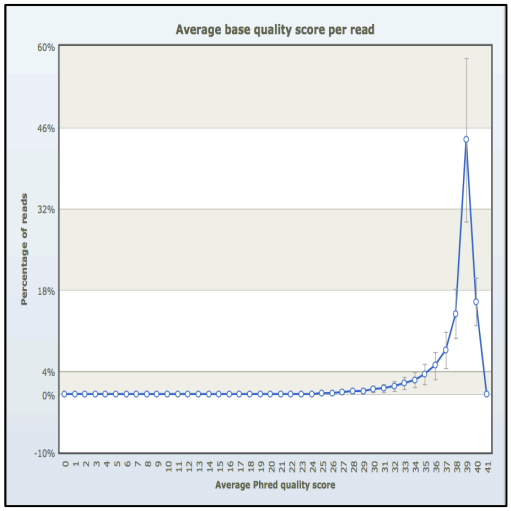


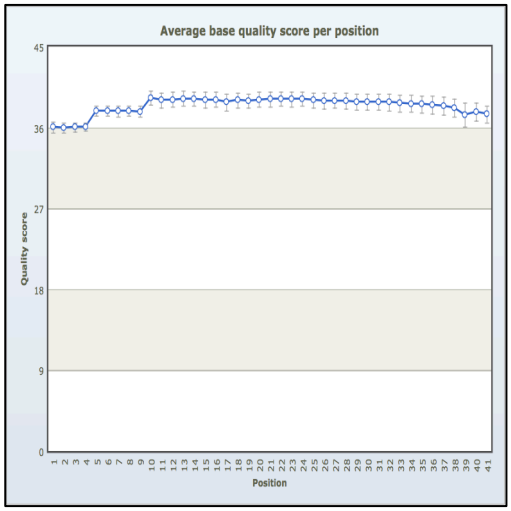
Figure 3- 3: Pre- alignment QA/QC- Bases Trimmed: (A): *Pre- alignment QA/QC of Average base quality score per position after Trimming of Bases.* This graph shows the average quality score of each base. The quality of all the bases was above 36 indicating that the trimming was successful. **(B):** *Pre- alignment QA/QC Average base quality score per read after Trimming of Bases.* This graph shows the quality score per read. This indicates that more than 100% of the reads have a Phred score higher than 35 with the majority of reads having a Phred score of 39.

Furthermore, *seed length* was changed from 28 to 16 because, again, the length of microRNAs is smaller than the default settings. Finally, the *alignments reported per read* was increased from 1 to 3, which increased the number of valid alignments reported. Bowtie aligner was selected because it is the most suitable for short, high quality reads with high similarity and highly unique alignments. A post- alignment QA/QC report was generated **(Figures 3- 4A, 3- 4B, 3- 4C, 3- 4D)** and the quality of the samples remained the same after alignment. The alignment per read **(Figure 3- 4C)** shows that the majority of the reads lined up 3 times as the parameter was set and that the rest lined up more than twice. This allows for confidence in the reported miR profile. Furthermore, the alignment percent **(Figure 3- 4D)** was extremely high with an average of 90% lining up to the miR base, which indicated that the samples did not contain notable contamination.

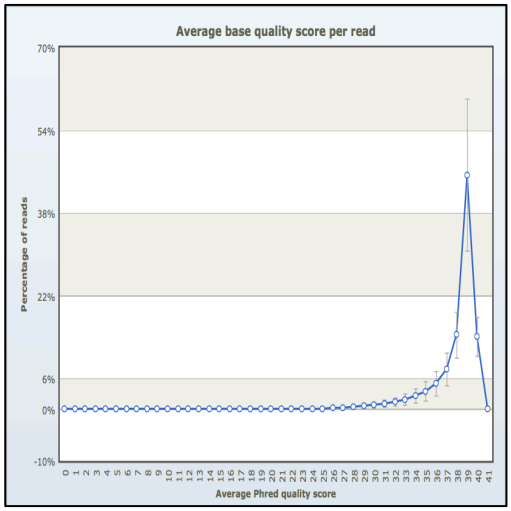
The data was then quantified to the transcriptome using Partek E/M, which reports the number of reads for each miR that was aligned to the transcriptome. This provides the user with a preliminary idea of what the miR profile looks like.

Differential expression analysis was then conducted using edgeR in order to normalize the raw sequencing reads. This step allows for the minimization of technical bias and ensures that the differences in miR expression are truly due to biological difference. As in any study, various steps introduce bias to the results. In the case of RNA sequencing the main concern is gene length, sequencing depth, and RNA composition[35].

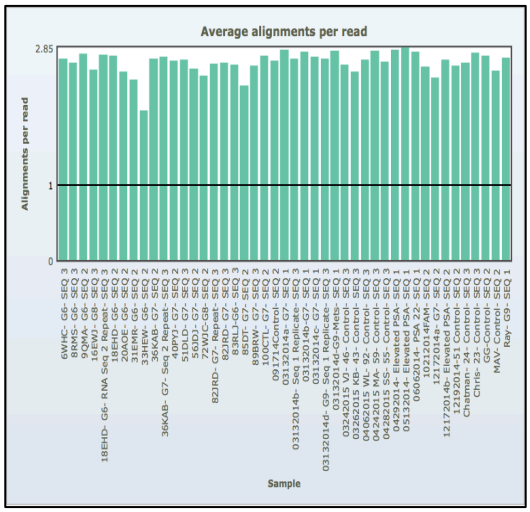
A.



B.



C.



D.

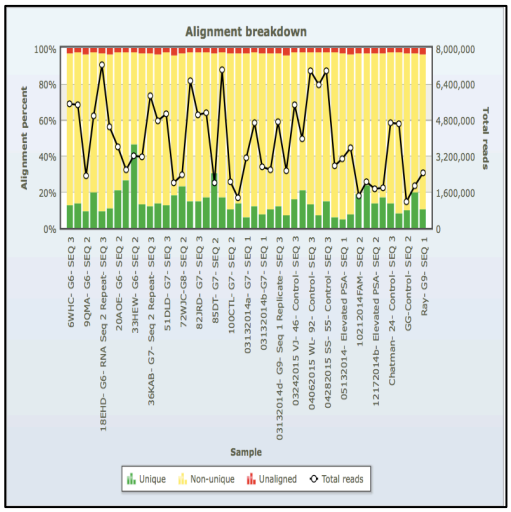


Figure 3- 4: Post- alignment QA/QC: **(A):** *Post- alignment QA/QC of Average base quality score per position after Trimming of Bases.* The quality of all the bases was above 36 indicating that the trimming was successful. **(B):** *Post- alignment QA/QC Average base quality score per read after Trimming of Bases.* This indicates that

more than 100% of the reads have a Phred score higher than 37 with the majority of reads having a Phred score of 39. **(C):** *Average Alignments per Read.* Majority of samples aligned between 2.5 to 3 times. **(D):** *Alignment Breakdown.* All samples had a high alignment percentage.

The most commonly utilized normalization methods for RNA sequencing is Reads Per Kilobase Per Million Mapped Reads (RPKM) and total counts. The RPKM method was first proposed by Mortazavi et al. (2008) and operates on the notion that the number of reads of a certain gene tends to be proportional to the length of the gene[36]. On the other hand, the total counts method divides the raw reads of each transcript in a sample by the corresponding total number of reads in the sample which adjusts for the library size[37].

However, these normalization methods have their limitations. Both of these methods only take the length of the gene and its expression level into account while excluding the effect of RNA composition, which also has a great impact on downstream differential expression analysis. In any given sample, highly expressed genes may cause an under representation of lower expressed genes leading to a false positive detection of dysregulated genes. Furthermore, microRNAs all fall within the 18- 22 nt length and therefore the use of gene length is not effective to normalize miRNA deep sequencing data. Thus, a new normalization method that places emphasis on RNA composition rather than length was necessary.

For this reason, Trimmed Mean of *M*-values (TMM) proposed by Robinson et al. (2012), which normalizes by considering RNA composition, was utilized instead[36]. This method relies on the hypothesis that most genes are in fact not differentially express. The edgeR software, the statistical component of R, operates using the TMM normalization method and was used for all differential expression analysis[38].

This generated 550 miRs with their associated p-values and false detection rates or false discovery rates (FDR) **(Table 3-3)**. A significant p value was any value at or below the standard value of 0.05. A value equal to or less than that standard cutoff suggests that the null hypothesis

(meaning there is no real dysregulation in the miR profile between controls and patients) can be rejected BUT does not mean that the dyregulation observed is actually due to biological variation. The most significant aspect of an extremely low p value is that it controls for Type 1 errors, which are the incorrect rejection of a true null hypotheses or false positives. The FDR further helps with minimizing Type 1 errors by controlling the proportion of rejected null hypotheses that were incorrect. Since there were 550 miRs expressed at appreciable values, a FDR value of 0.2 was selected meaning that ≈110 out of 550 miRs would be expected to produce false positives. With the parameters set as stringently as possible, this produced 26 miRs to investigate further for dyregulation **(Table 3-3)**.

The resulting data was used to determine which miRs to pursue for the normalization and dysregulation studies. An analysis of these results will be discussed in the following chapters.

Table 3- 3: Sample Log List Generated from RNA Sequencing Results

MicroRNA	P Value	FDR
hsa-miR-5582-3p	2.36E-06	0.001298179
hsa-miR-32-5p	1.23E-05	0.00337548
hsa-miR-500b-3p	0.000572225	0.10471709
hsa-miR-329-3p	0.001109288	0.131868458
hsa-miR-487b-3p	0.001200988	0.131868458
hsa-miR-454-3p	0.00150438	0.137650814
hsa-miR-204-5p	0.001930515	0.151407567
hsa-miR-20a-5p	0.002972257	0.167354157
hsa-miR-127-3p	0.003157203	0.167354157
hsa-miR-543	0.00347709	0.167354157
hsa-miR-5001-3p	0.004038855	0.167354157
hsa-miR-181a-3p	0.004057331	0.167354157
hsa-miR-144-3p	0.004134586	0.167354157
hsa-miR-660-3p	0.004346684	0.167354157
hsa-miR-374a-5p	0.004572518	0.167354157
hsa-miR-128-3p	0.005373932	0.184393041
hsa-miR-654-5p	0.006803665	0.206237737
hsa-miR-17-5p	0.006906872	0.206237737
hsa-miR-4676-3p	0.007747569	0.206237737
hsa-miR-20b-5p	0.007983965	0.206237737
hsa-miR-18b-5p	0.008218656	0.206237737
hsa-miR-134-5p	0.008416145	0.206237737
hsa-let-7f-2-3p	0.00944494	0.206237737
hsa-miR-642a-5p	0.009733771	0.206237737
hsa-miR-942-5p	0.010041576	0.206237737
hsa-miR-4433a-3p	0.010133679	0.206237737

Chapter 4

Normalization

Introduction:

After analysis of the RNA sequencing results, it became evident that a normalization method needed to be developed to validate miR dysregulation detected in qRT-PCR assays as due to biological variation or just technical error. Variation within a data set can be due to "true" biological variation, i.e. caused by changes within the organism, which experimenters wish to document or due to technical variations that are caused by experimental errors. During the course of any study, many systemic and random errors can accumulate, which add to the variation in the results. Systemic errors affect all the samples in a data set in the same way and cause similar deviations from the real value causing predictable and proportional differences. Random errors, however, as the name suggests, are much more difficult to eliminate because they instill unpredictable fluctuations in the results effecting the precision. Such errors can arise from the experimenter or instabilities in the apparatus.

For this reason, normalization techniques are imperative to eliminate the bias that arises from these errors and to ensure that the observed variation is in fact due to "true" biological variation. The best way to normalize is to use an endogenous control. Such controls should be present within the organism as part of the miR profile and are stably expressed across both patients and controls.

One of the major disconcerting aspects of many microRNA studies is the lack of a reliable normalizer. Many studies still rely on the use of members of the RNU family, also known as snoRNAs, which are small nucleolar RNAs. However, these non- coding RNAs have been found to be unreliable endogenous controls when dealing with whole blood, serum, and other body fluids because they are located within cells and are not actively secreted into the

body fluids unless left behind as degradation products from cellular debris. Moreover, one study analyzed breast cancer paraffin embedded samples and found that snoRNAs were actually downregulated[39]. Another limiting factor is that normalization candidates can vary between different body fluids and diseases, making it even more difficult to find a universal normalization technique. Thus normalization standards had to first be determined before confirmatory miR dysregulation studies could be pursued.

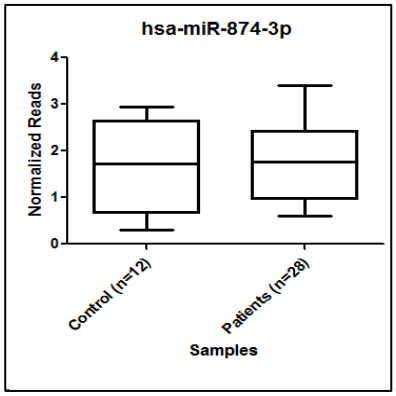
For the purposes of this study, the RNA sequencing data was used to determine which miRs were stably expressed across all the samples. Once normalizers were identified, several factors were taken into consideration to determine which miRs should be analyzed further using qRT-PCR. The first factor was the standard deviation between the normalized reads. A small standard deviation was desired to ensure that the normalized miRs exhibited constant expression amongst all the samples. Second, the abundance was taken into consideration. Very low abundance miRs would yield high CT reflecting lower concentrations in the sample, therefore proving to be unreliable normalized candidates. On the other hand, miRs that were expressed too highly could wipe out dysregulation differences and instill bias in the results. Finally, the coefficient variance was calculated (standard deviation divided by the mean) to standardize the variability between sample populations.

Results:

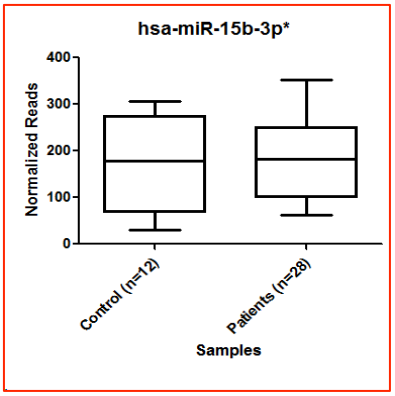
After analysis of the RNA sequencing results, several candidates showed convincing potential as normalizers including hsa-miR-874-3p, hsa-miR-15b-3p, hsa-miR-6783-3p, hsa-miR-4511, hsa-miR-6842-3p, hsa-miR-139-3p, hsa-miR-1468-5p, hsa-miR-92a-1-5p, hsa-miR-202-5p, hsa-miR-197-3p, hsa-miR-671-3p, hsa-miR-933, hsa-miR-1248, hsa-miR-628-5p, hsa-miR-146a-5p, hsa-miR-92b-5p, hsa-miR-5193, hsa-miR-342-5p, hsa-miR-659-5p, hsa-miR-23a-3p, hsa-miR-423-3p, hsa-miR-3909, hsa-miR-296-5p, hsa-miR-340-3p, hsa-miR-769-5p, hsa-miR-1538. A box plot of the RNA sequencing data for each of these miRs in control and patient samples is shown in **Figure 4-1A through Figure 4-1Z**.

A review of the literature was also conducted to determine if there was any overlap between these potential miRs and the RNA sequencing results. Although some of these studies investigated miRs in different body fluids, it was important include a literature review to ensure that the best candidates were chosen for further investigation. Several miRs were suggested to be suitable normalization miRs are as follows; hsa-miR-130b in prostate tissue[40], hsa-miR-191 in the urine of prostate cancer patients and controls[41], hsa-miR-103 in T-cell acute lymphoblastic leukemia (T-ALL) samples[42], and hsa-miR-93 in normal and cancerous human solid tissues[43]. Box plots of their analysis in the RNA sequencing data are shown in **Figure 4-2A through Figure 4-2D**.

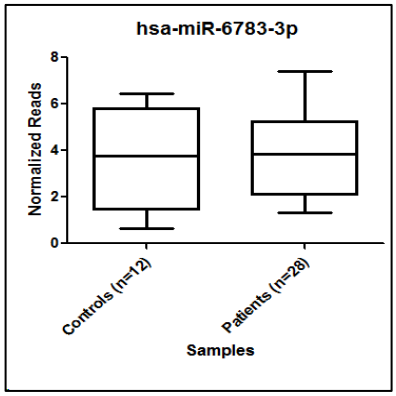
A.


hsa-miR-874-3p
Normalized Reads
0
1
2
3
4
Control (n=12)
Patients (n=28)
Samples

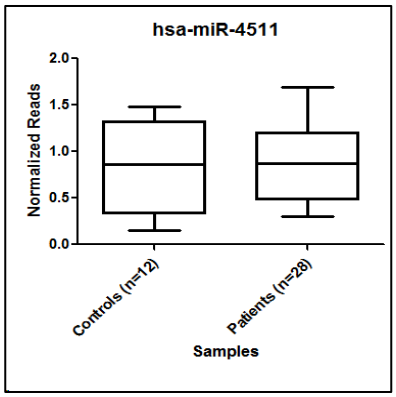
B.


hsa-miR-15b-3p*
Normalized Reads
0
100
200
300
400
Control (n=12)
Patients (n=28)
Samples

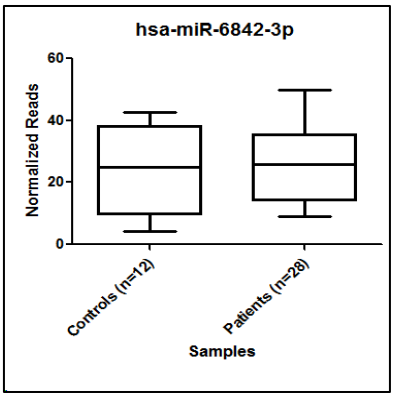
C.


hsa-miR-6783-3p
Normalized Reads
0
2
4
6
8
Controls (n=12)
Patients (n=28)
Samples

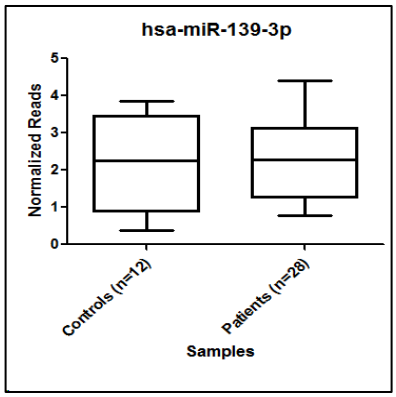
D.


hsa-miR-4511
Normalized Reads
0.0
0.5
1.0
1.5
2.0
Controls (n=12)
Patients (n=28)
Samples

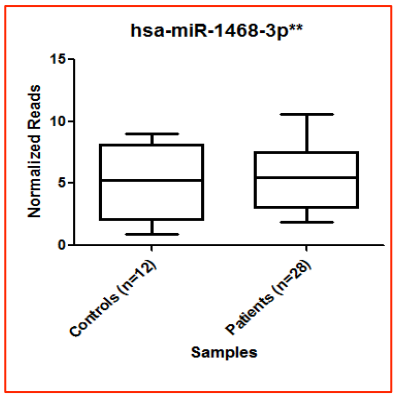
E.


hsa-miR-6842-3p
Normalized Reads
0
20
40
60
Controls (n=12)
Patients (n=28)
Samples

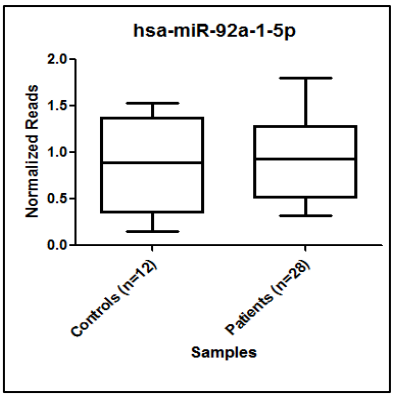
F.


hsa-miR-139-3p
Normalized Reads
0
1
2
3
4
5
Controls (n=12)
Patients (n=28)
Samples

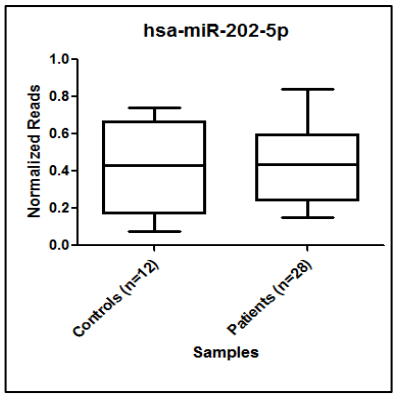
G.


hsa-miR-1468-3p**
Normalized Reads
0
5
10
15
Controls (n=12)
Patients (n=28)
Samples

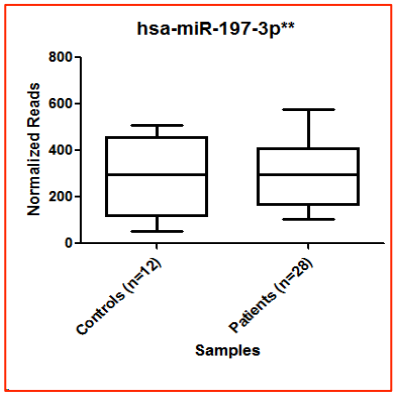
H.


hsa-miR-92a-1-5p
Normalized Reads
0.0
0.5
1.0
1.5
2.0
Controls (n=12)
Patients (n=28)
Samples

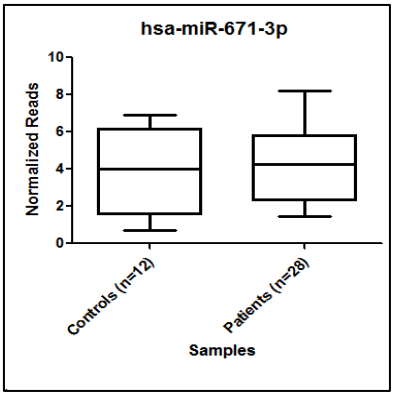
I.


hsa-miR-202-5p
Normalized Reads
0.0
0.2
0.4
0.6
0.8
1.0
Controls (n=12)
Patients (n=28)
Samples

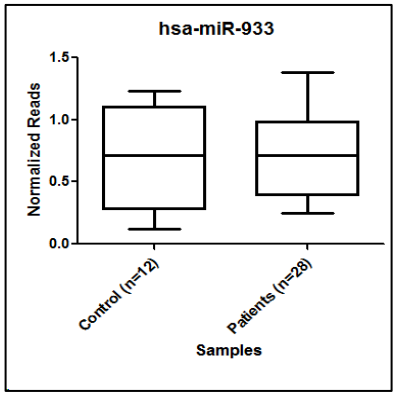
J.


hsa-miR-197-3p**
Normalized Reads
800
600
400
200
0
Controls (n=12)
Patients (n=28)
Samples

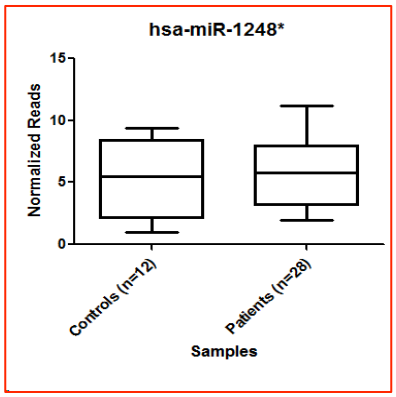
K.


hsa-miR-671-3p
Normalized Reads
10
8
6
4
2
0
Controls (n=12)
Patients (n=28)
Samples

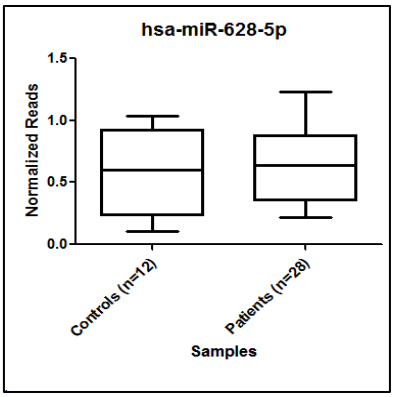
L.


hsa-miR-933
Normalized Reads
1.5
1.0
0.5
0.0
Control (n=12)
Patients (n=28)
Samples

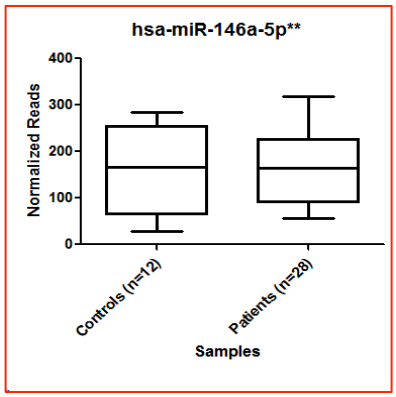
M.


hsa-miR-1248*
Normalized Reads
15
10
5
0
Controls (n=12)
Patients (n=28)
Samples

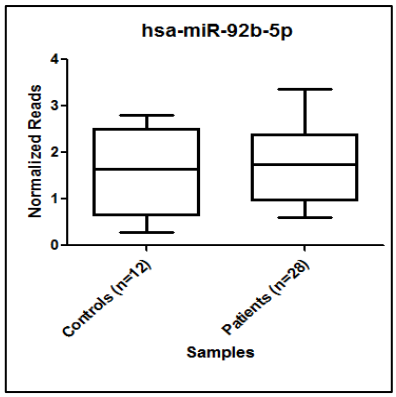
N.


hsa-miR-628-5p
Normalized Reads
1.5
1.0
0.5
0.0
Controls (n=12)
Patients (n=28)
Samples

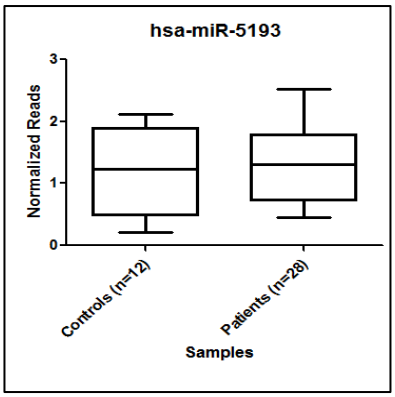
O.


hsa-miR-146a-5p**
Normalized Reads
400
300
200
100
0
Controls (n=12)
Patients (n=28)
Samples

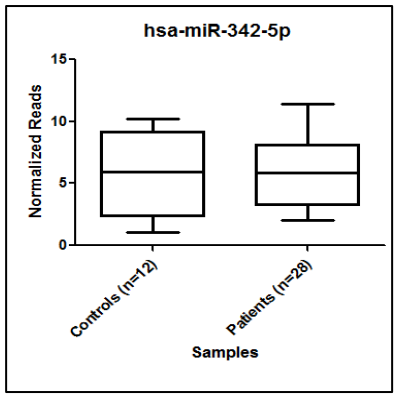
P.


hsa-miR-92b-5p
Normalized Reads
4
3
2
1
0
Controls (n=12)
Patients (n=28)
Samples

Q.


hsa-miR-5193
Normalized Reads
3
2
1
0
Controls (n=12)
Patients (n=28)
Samples

R.


hsa-miR-342-5p
Normalized Reads
15
10
5
0
Controls (n=12)
Patients (n=28)
Samples

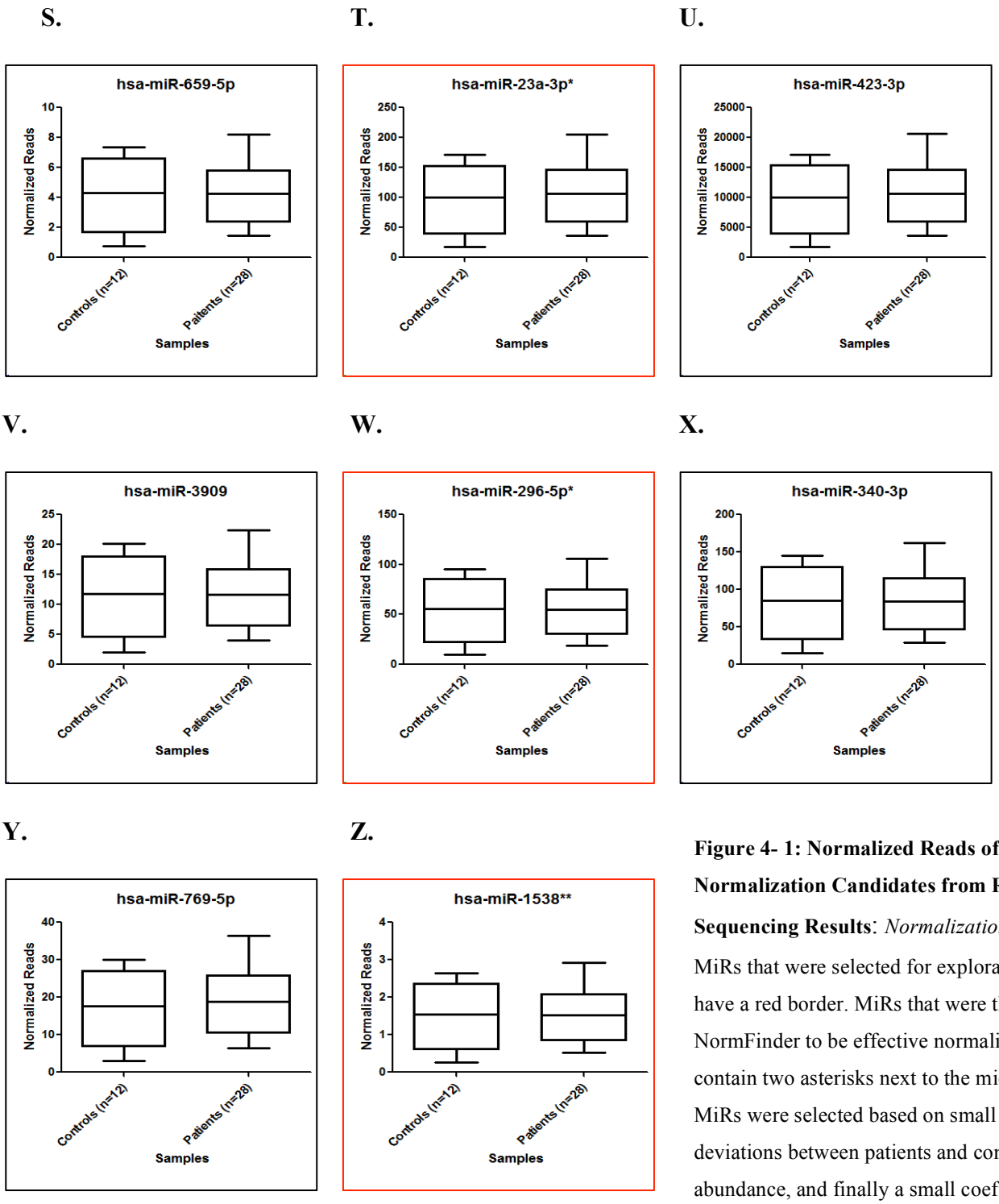
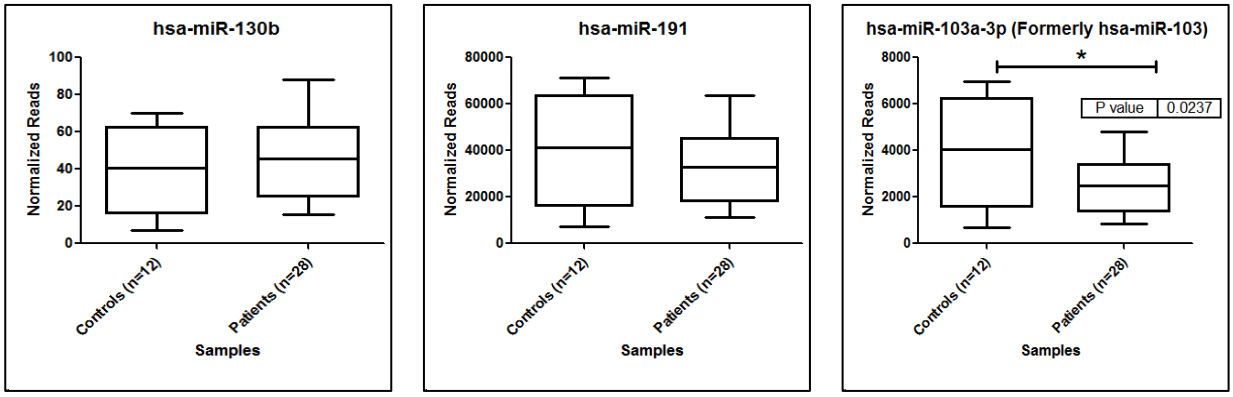


Figure 4- 1: Normalized Reads of Possible Normalization Candidates from RNA Sequencing Results: *Normalization Candidates:* MiRs that were selected for exploratory qRT-PCR have a red border. MiRs that were then selected by NormFinder to be effective normalization standards contain two asterisks next to the microRNA name. MiRs were selected based on small standard deviations between patients and controls, their abundance, and finally a small coefficient variance.

A. **B.** **C.**



D.

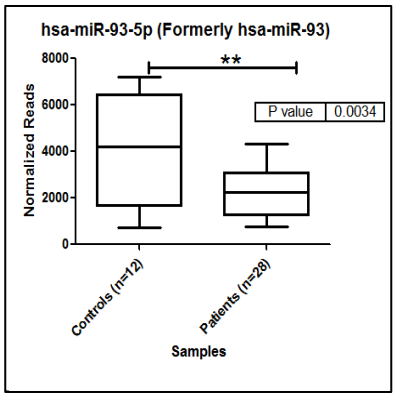


Figure 4- 2: Normalized Reads of Normalization Candidates from Literature Review: *Candidates suggested by the literature as normalization candidates.* **(A):** MiR-130b was not selected as a normalization candidate due to the difference in the means between patient and control samples. **(B):** MiR- 191 was found to be extremely abundant in the samples and was not selected for normalization for fear that it would wipe out dyregulation. **(C) and (D):** These miRs were found to be extremely dysregulated displayed extremely low p values associated with them. As a result they were not selected as normalization candidates.

As can be seen in the graphs in **Figures 4-1 and 4-2,** some of the candidates were too abundantly expressed, were expressed at very low levels, or were too variable among the sample population. Moreover two of the miRs suggested in the literature **(Figures 4-2C and 4-2D)** showed dyregulation and even presented extremely low p values as shown in their respective graphs. Therefore, of the 30 miRs suggested via RNA sequencing results, 8 were selected for exploratory qRT- PCR **(Figure 4-1- figures with one asterisk or two asterisks were selected for exploratory PCR)**.

The results of the CT values generated for each selected miR in qRT- PCR data are shown in **Figure 4-3**. Interestingly, most of the normalization candidates produced extremely small standard deviations between samples within the same group or different groups **(Table 4-1)**. Small standard deviations are desired for normalization candidates. NormFinder analyzes intra- and intergroup variations to determine which candidates are best suited. Although miR-197 has a higher standard deviation compared to two other miRs, its standard deviation is consistent between controls and patients. These standard deviations were smaller than those observed in the RNA sequencing results, although all these selected miRs all proved to be promising.

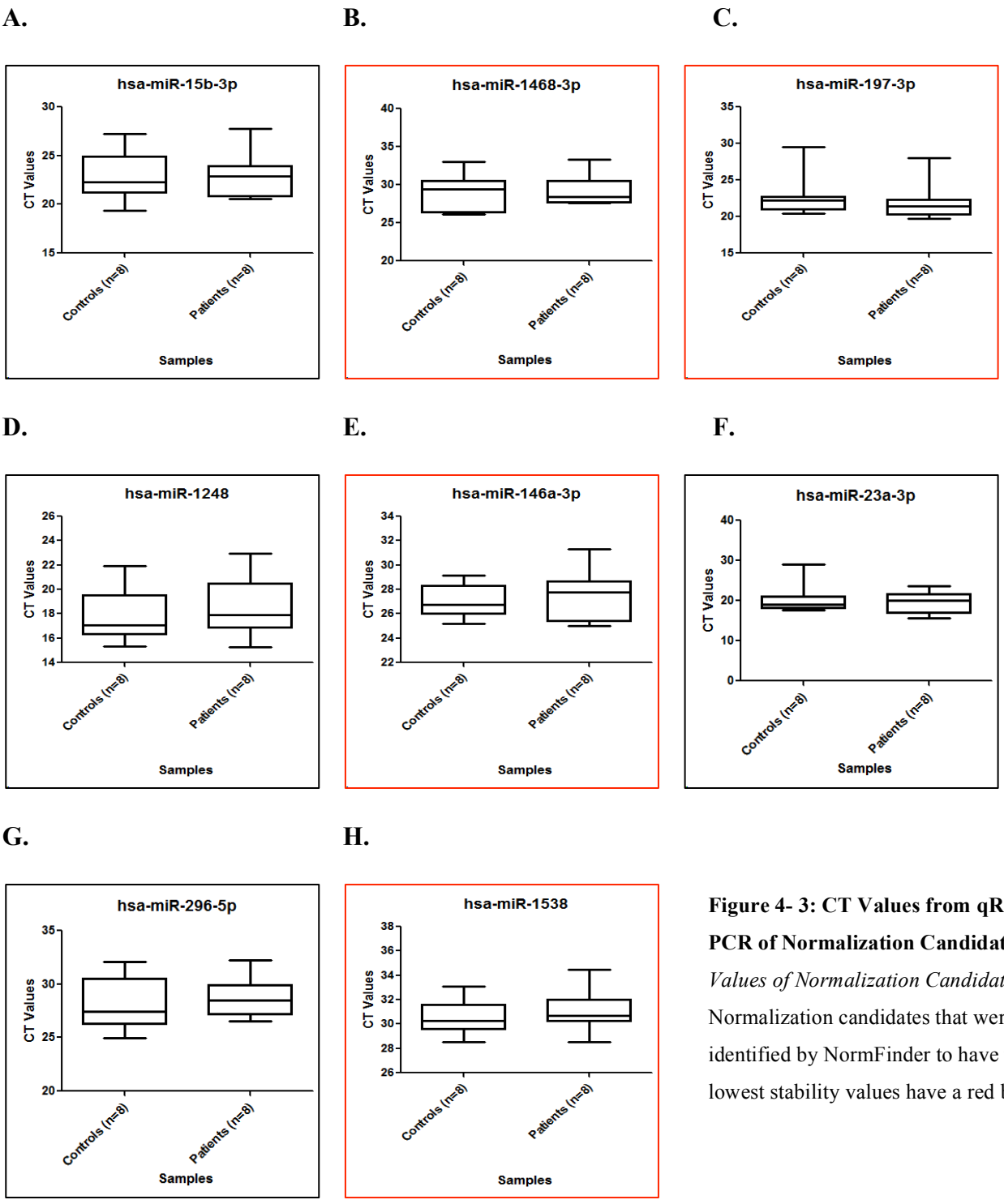


Figure 4- 3: CT Values from qRT-PCR of Normalization Candidates: *CT Values of Normalization Candidates*: Normalization candidates that were identified by NormFinder to have the lowest stability values have a red border.

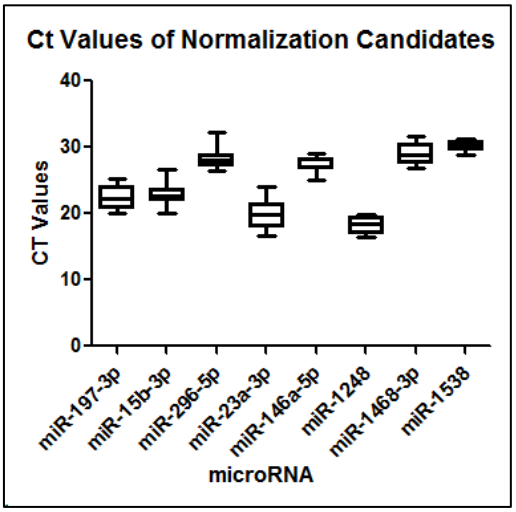


Figure 4- 4: CT Values from qRT-PCR of Normalization Candidates: *Graph depicting the collective CT Values of all the normalization candidates:* As can be seen miR-1538, miR- 1468-3p, and miR-146a-5p all exhibited small standard deviations as well as similar CT values. MiR-1248 and miR-197 also displayed small standard deviation although lower CT values.

Table 4- 1: Standard Deviation of Normalization miRs

MiR	Std between Controls	Std between Patients	Std between Controls and Patients
MiR 197-3p	2.896500338	2.644989917	2.708628336
MiR 15b-3p	2.544328016	2.370826007	2.377684418
MiR 296-5p	2.467039537	1.840413884	2.128742552
MiR 23a-3p	3.776250299	2.837212083	3.249625057
MiR 146a- 5p	1.344533392	2.078563461	1.713997946
MiR 1248	2.168850477	2.497720751	2.288752129
MiR 1468-3p	2.438152474	2.070897156	2.186690811
MiR 1538	1.273392193	1.544603799	1.382686823

NormFinder:

After exploratory PCR was conducted on the 8 selected candidates, the statistical program NormFinder was then utilized to determine which combinations of these miRs would provide the best normalization standards. NormFinder was chosen as opposed to other available programs that are available because the authors use an algorithm that is NOT susceptible to preferring miRs that are co-regulated. It was felt that such a preference could further bias our results. This program uses a linear model that takes into account intragroup and intergroup variation. This means that observed fluctuations that are seen between individuals in the same group are compared to fluctuations seen between patients and controls.

Based on this analysis a stability value is allocated to each miR. A small stability value is associated with the best normalization candidate because it indicates small intra- and intergroup variability **(Figure 4-5)**. The bolded numbers that are in blue represent the four best values and the miRs that were selected as the normalization panel for this study. The program also provides the user with the best combination of two miRs and the composite stability value of those two miRs **(Table 4-2)**. A lower stability value for a combination of two miRs ensures that smaller variations are seen between the two groups. The program does not allow for the analysis of any computation of miRs greater than two. Thus, it was not possible to calculate a stability value for all four miRs together, but since the stability value of a single miR (miR-146a) went from 0.009 to 0.008 with the addition of miR-1538, it is projected that a combined stability of all four would be extremely favorable. Therefore, for the purposes of this study, a geometric mean of the four best miRs determine by NormFinder were determined to be the most reliable.

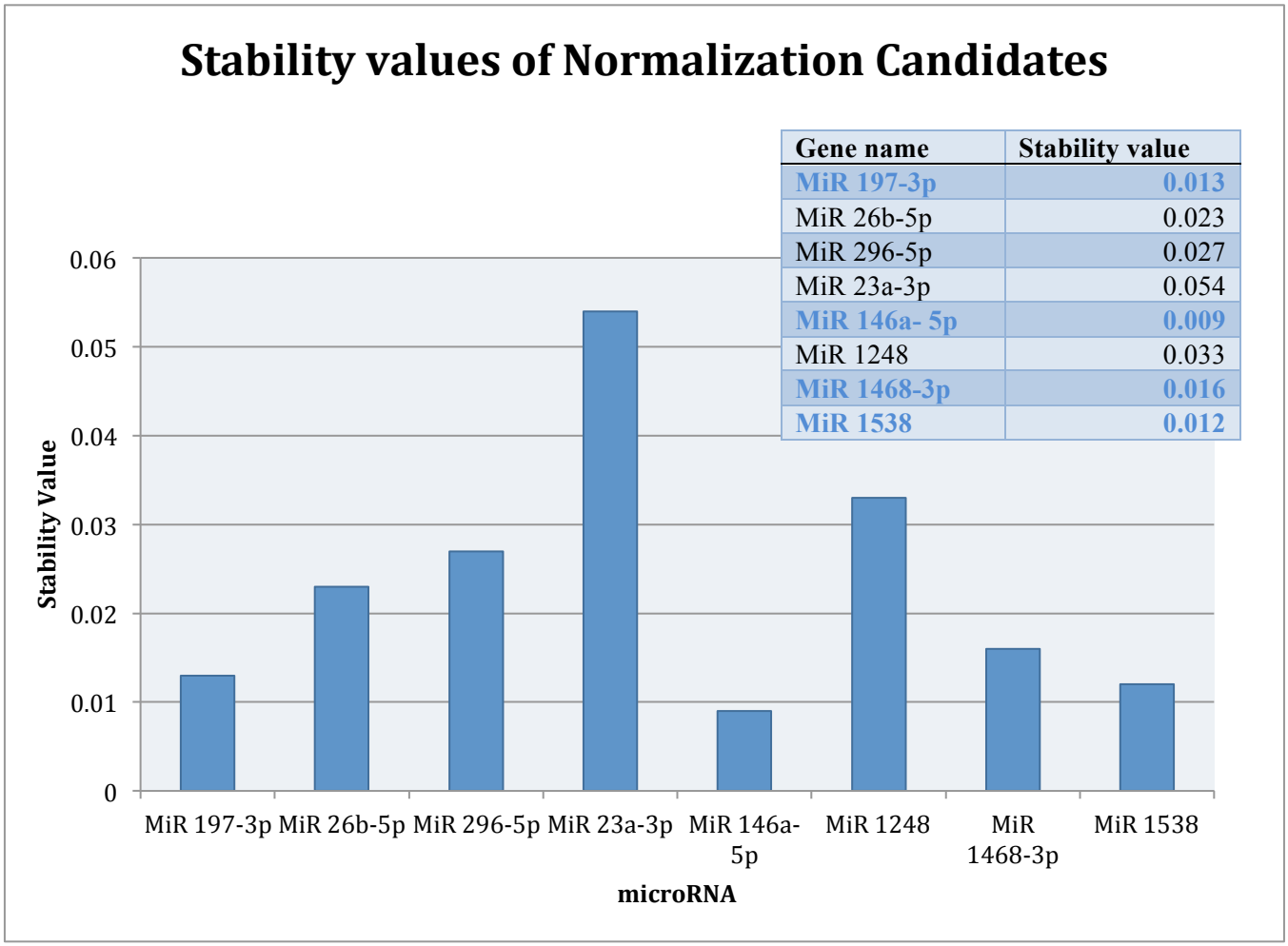


Figure 4- 5: Stability Value Generated by NormFinder: *Stability Values and their corresponding miRs:* A list of the stability values of the eight microRNAs explored via qRT-PCR. The ones that are bolded and are in blue represent the four best values and the miRs that were selected as the normalization panel for this study.

Table 4- 2: Top miRs Selected by NormFinder

Best gene	**MiR 146a- 5p**
Stability value	0.009
Best combination of two genes	**MiR 146a- 5p and MiR 1538**
Stability value for best combination of two genes	0.008

Chapter 5

Dysregulated miRs Results

Analysis:

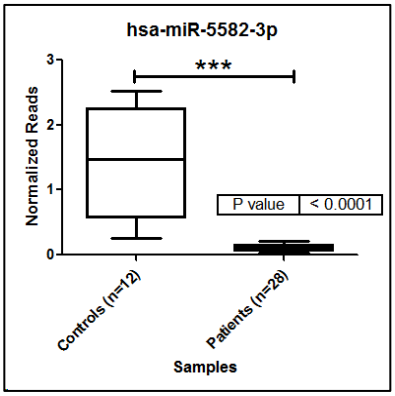
The RNA sequencing data of blood revealed that there were potentially 26 dysregulated miRs that yielded acceptable p- values and FDR values (<0.05 and 0.2 respectively) **(Table 3-1 and Figure 5-1A through Figure 5-1Z)**. Analysis of the box plots showed that the best candidates to pursue for further validation were the first nine miRs with the lowest p- values and lowest FDRs **(Figure 5-1A- Figure 5-1I)**. However, the primers for hsa-miR-5582-3p and hsa-miR-500b-3p could not be obtained by Quanta in time for the purposes of this study and will therefore be analyzed at a later time **(Figure 5-1A and Figure 5-1C)**.

Consequently, that left four upregulated miRs and three downregulated miRs. Hsa-miR-127-3p, hsa-miR-204-5p, hsa-miR-329-3p and hsa-miR-487b-3p where shown to be upregulated in the RNA seq data and hsa-miR-20a-5p, hsa-miR-32-5p and hsa-miR-454-3p was observed to be downregulated.

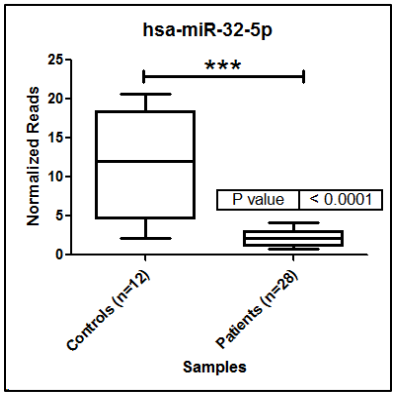
Results:

qRT-PCR assays confirmed these findings **(Figure 5-2A through Figure 5-2N)**. The box plots as well as their corresponding dot plots are shown along with their associated p-values. The plots were produced by subtracting the raw CT value of the miR being studied by the geometric mean of the four selected normalizers. Both plots were shown because the box plot effectively displays the standard deviation of the samples within each group, whereas if outliers are present they can be more clearly noted in the dot plot.

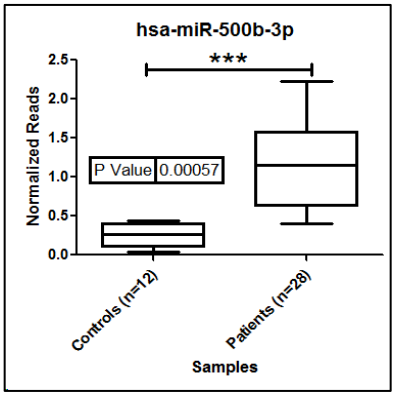
A.


hsa-miR-5582-3p
***
Normalized Reads
P value < 0.0001
Controls (n=12)
Patients (n=28)
Samples

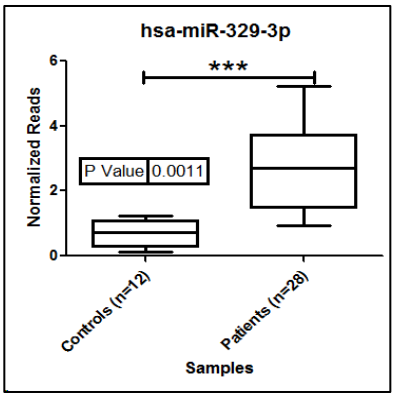
B.


hsa-miR-32-5p
***
Normalized Reads
P value < 0.0001
Controls (n=12)
Patients (n=28)
Samples

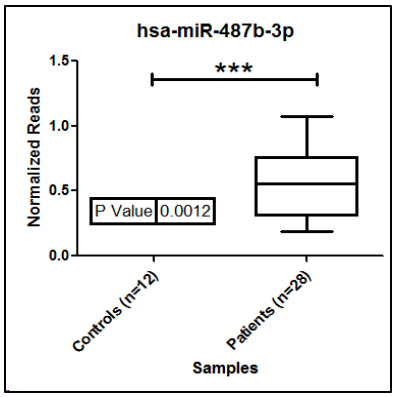
C.


hsa-miR-500b-3p
***
Normalized Reads
P Value 0.00057
Controls (n=12)
Patients (n=28)
Samples

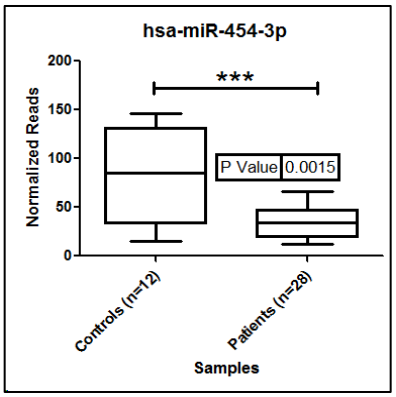
D.


hsa-miR-329-3p
***
Normalized Reads
P Value 0.0011
Controls (n=12)
Patients (n=28)
Samples

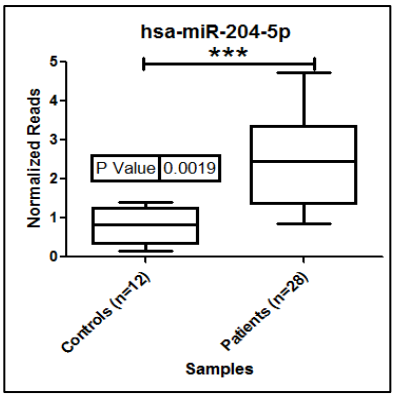
E.


hsa-miR-487b-3p
***
Normalized Reads
P Value 0.0012
Controls (n=12)
Patients (n=28)
Samples

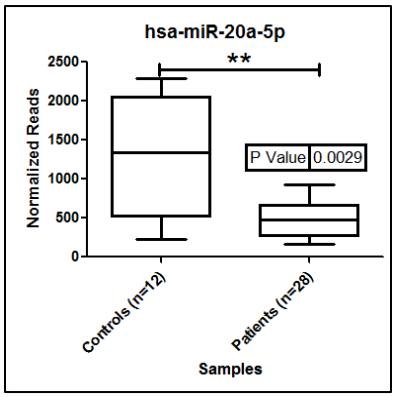
F.


hsa-miR-454-3p
***
Normalized Reads
P Value 0.0015
Controls (n=12)
Patients (n=28)
Samples

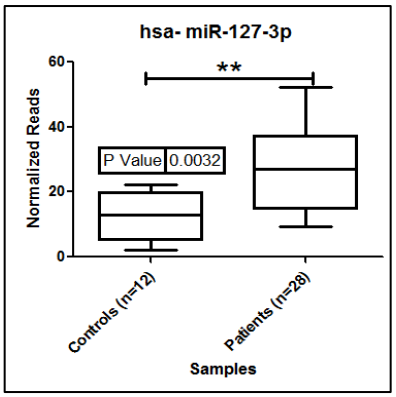
G.


hsa-miR-204-5p
***
Normalized Reads
P Value 0.0019
Controls (n=12)
Patients (n=28)
Samples

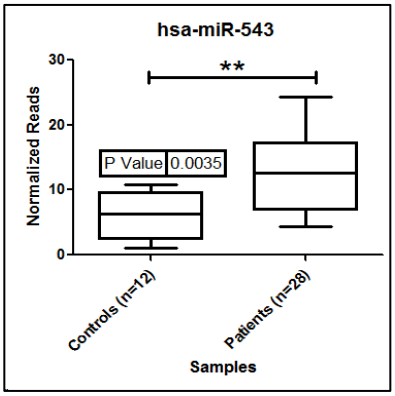
H.


hsa-miR-20a-5p
**
Normalized Reads
P Value 0.0029
Controls (n=12)
Patients (n=28)
Samples

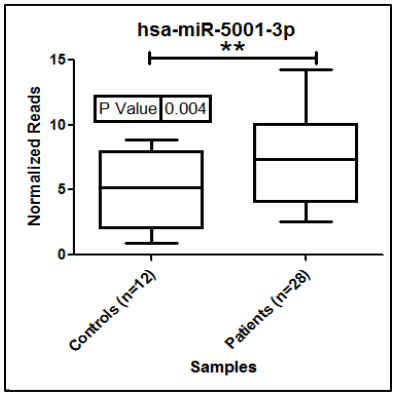
I.


hsa- miR-127-3p
**
Normalized Reads
P Value 0.0032
Controls (n=12)
Patients (n=28)
Samples

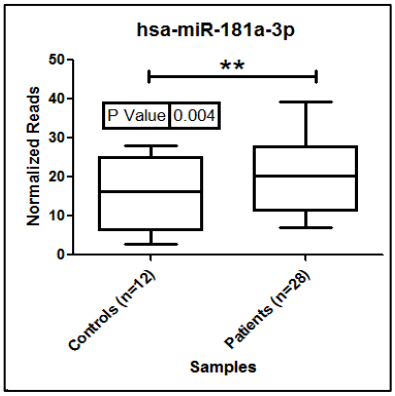
J.


hsa-miR-543
**
P Value 0.0035
Normalized Reads
30
20
10
0
Controls (n=12)
Patients (n=28)
Samples

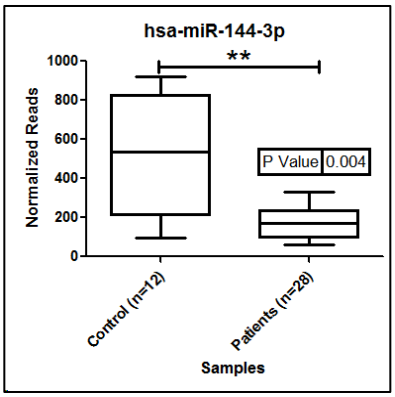
K.


hsa-miR-5001-3p
**
P Value 0.004
Normalized Reads
15
10
5
0
Controls (n=12)
Patients (n=28)
Samples

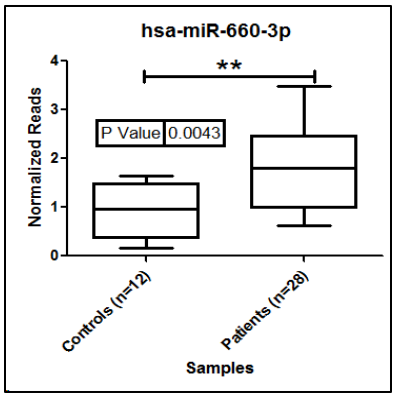
L.


hsa-miR-181a-3p
**
P Value 0.004
Normalized Reads
50
40
30
20
10
0
Controls (n=12)
Patients (n=28)
Samples

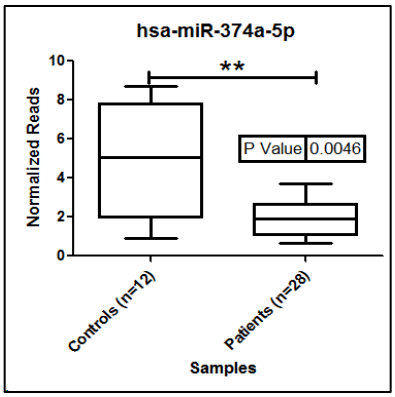
M.


hsa-miR-144-3p
**
P Value 0.004
Normalized Reads
1000
800
600
400
200
0
Control (n=12)
Patients (n=28)
Samples

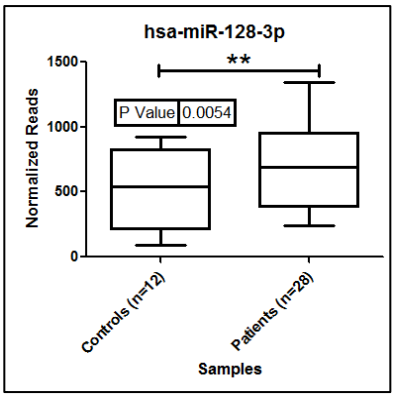
N.


hsa-miR-660-3p
**
P Value 0.0043
Normalized Reads
4
3
2
1
0
Controls (n=12)
Patients (n=28)
Samples

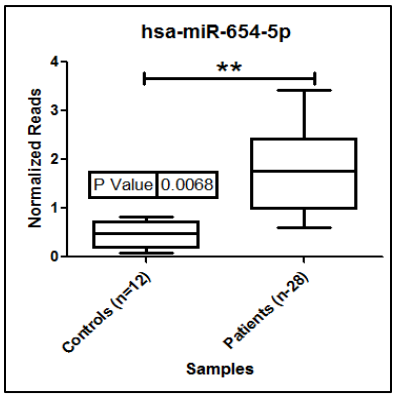
O.


hsa-miR-374a-5p
**
P Value 0.0046
Normalized Reads
10
8
6
4
2
0
Controls (n=12)
Patients (n=28)
Samples

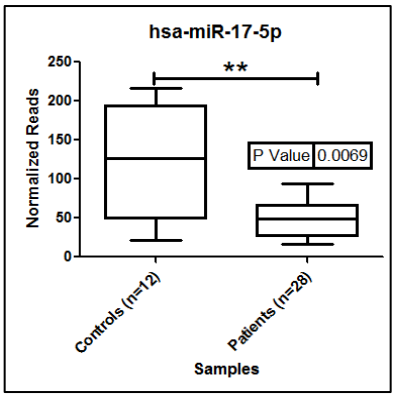
P.


hsa-miR-128-3p
**
P Value 0.0054
Normalized Reads
1500
1000
500
0
Controls (n=12)
Patients (n=28)
Samples

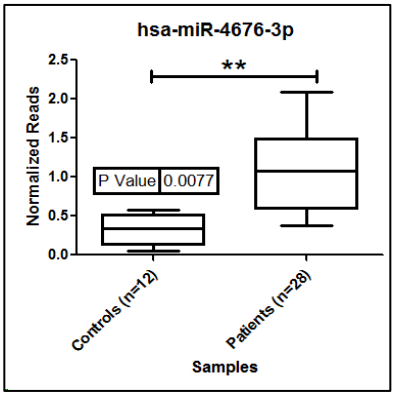
Q.


hsa-miR-654-5p
**
P Value 0.0068
Normalized Reads
4
3
2
1
0
Controls (n=12)
Patients (n=28)
Samples

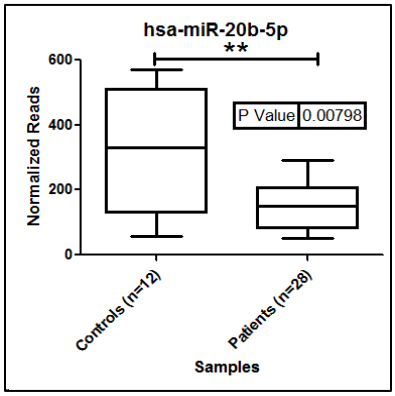
R.


hsa-miR-17-5p
**
P Value 0.0069
Normalized Reads
250
200
150
100
50
0
Controls (n=12)
Patients (n=28)
Samples

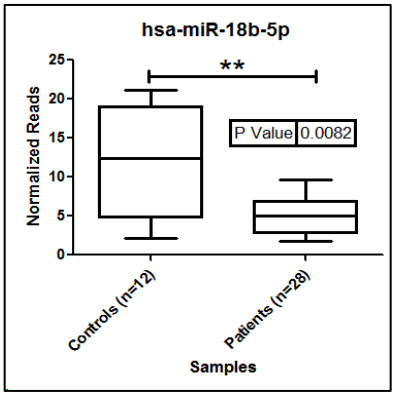
S.



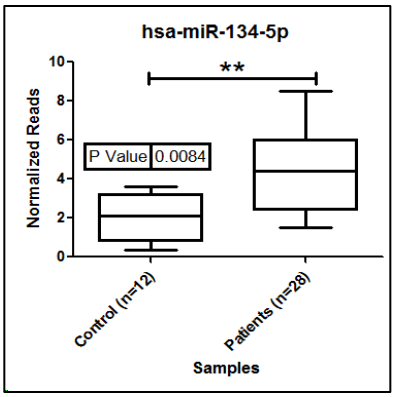
T.



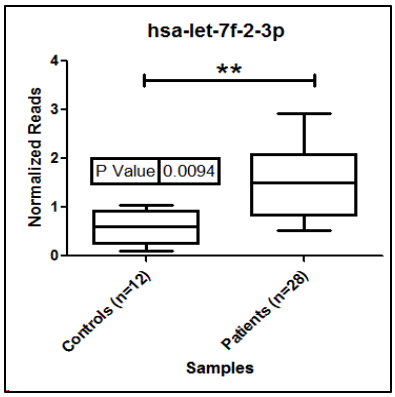
U.



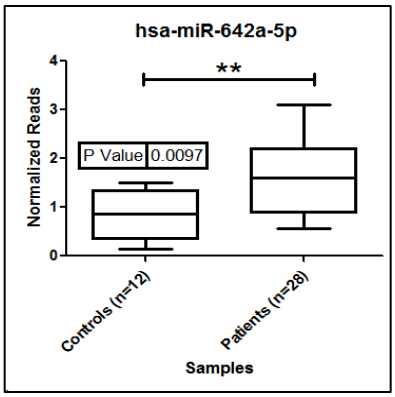
V.



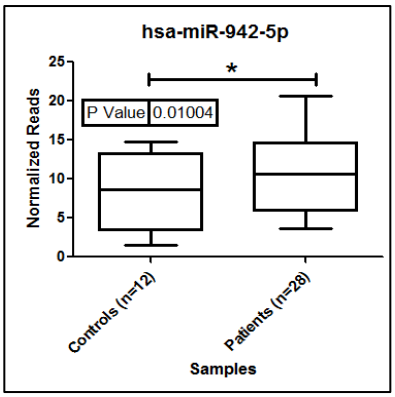
W.



X.



Y.



Z.

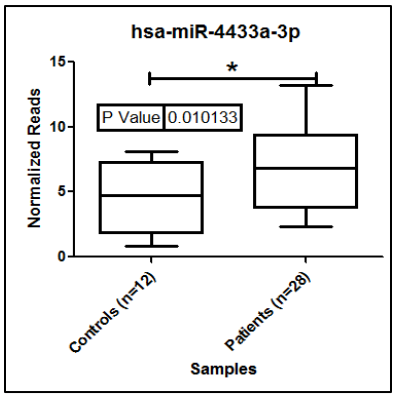
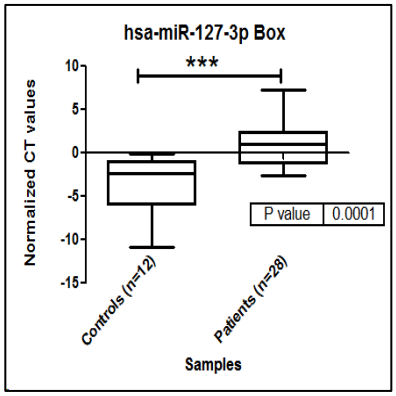
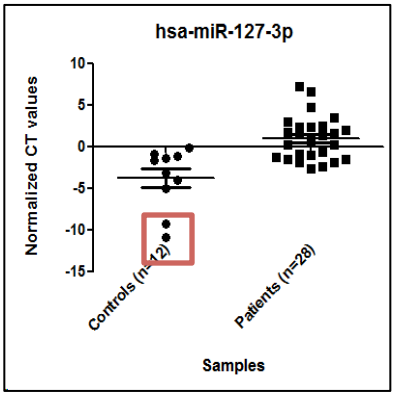


Figure 5-1: Normalized Reads of Dysregulated miRs: *Normalization Reads of Dysregulated Candidates*: Dysregulated candidates were selected based on p values, FDR values and limited overlap between patients and controls. Asterisks represent the p-value. Three asterisks are smaller p-values compared to the one asterisk, which has a p-value of less than 0.05 but higher than 0.01.

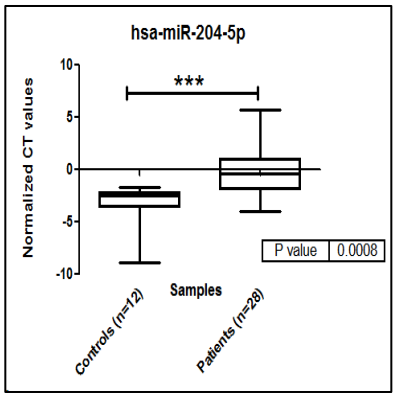
A.


hsa-miR-127-3p Box
***
Normalized CT values
10
5
0
-5
-10
-15
P value
0.0001
Controls (n=12)
Patients (n=28)
Samples

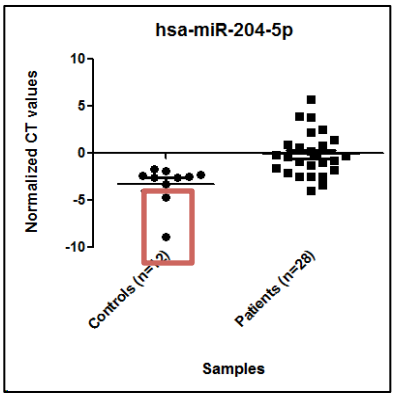
B.


hsa-miR-127-3p
Normalized CT values
10
5
0
-5
-10
-15
Controls (n=12)
Patients (n=28)
Samples

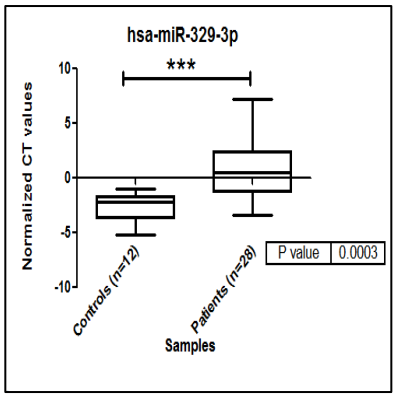
C.


hsa-miR-204-5p
***
Normalized CT values
10
5
0
-5
-10
P value
0.0008
Samples
Controls (n=12)
Patients (n=28)

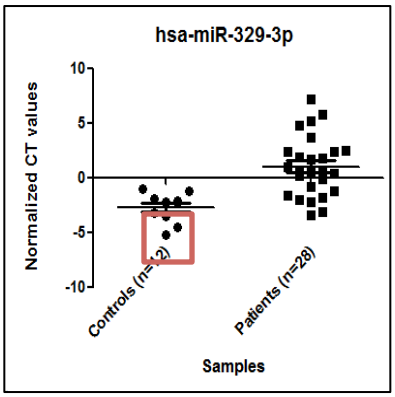
D.


hsa-miR-204-5p
Normalized CT values
10
5
0
-5
-10
Controls (n=12)
Patients (n=28)
Samples

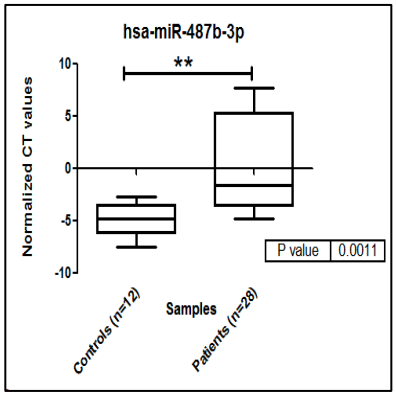
E.


hsa-miR-329-3p
***
Normalized CT values
10
5
0
-5
-10
P value
0.0003
Controls (n=12)
Patients (n=28)
Samples

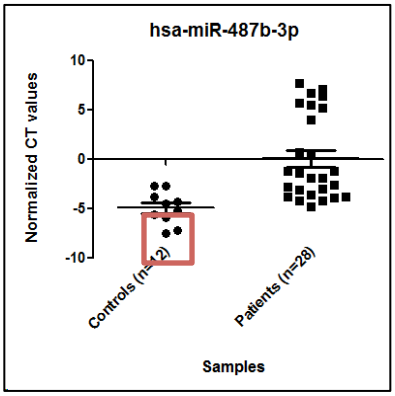
F.


hsa-miR-329-3p
Normalized CT values
10
5
0
-5
-10
Controls (n=12)
Patients (n=28)
Samples

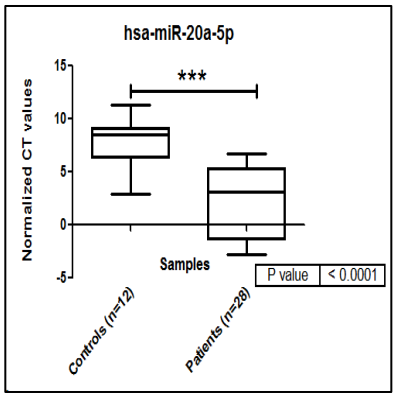
G.


hsa-miR-487b-3p
**
Normalized CT values
10
5
0
-5
-10
P value
0.0011
Samples
Controls (n=12)
Patients (n=28)

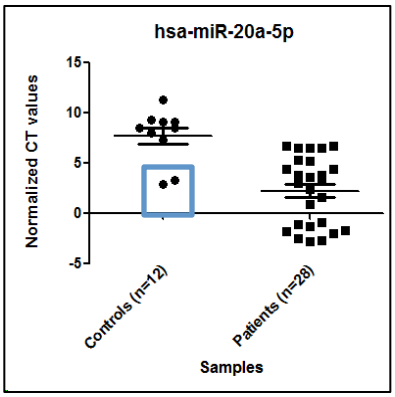
H.


hsa-miR-487b-3p
Normalized CT values
10
5
0
-5
-10
Controls (n=12)
Patients (n=28)
Samples

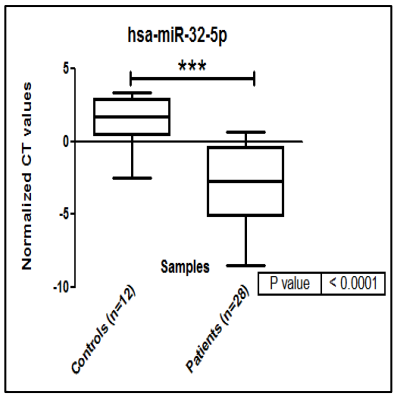
I.


hsa-miR-20a-5p
***
Normalized CT values
15
10
5
0
-5
Samples
P value
< 0.0001
Controls (n=12)
Patients (n=28)

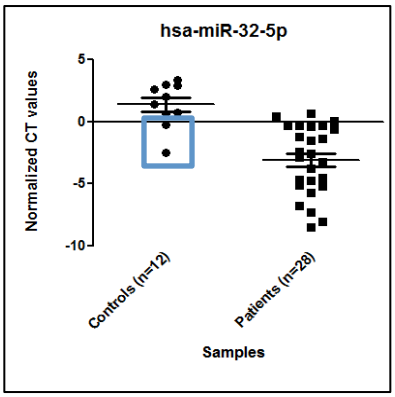
J.


hsa-miR-20a-5p
Normalized CT values
15
10
5
0
-5
Controls (n=12)
Patients (n=28)
Samples

K.


hsa-miR-32-5p
***
Normalized CT values
5
0
-5
-10
Samples
P value
< 0.0001
Controls (n=12)
Patients (n=28)

L.


hsa-miR-32-5p
Normalized CT values
5
0
-5
-10
Controls (n=12)
Patients (n=28)
Samples

M. **N.**

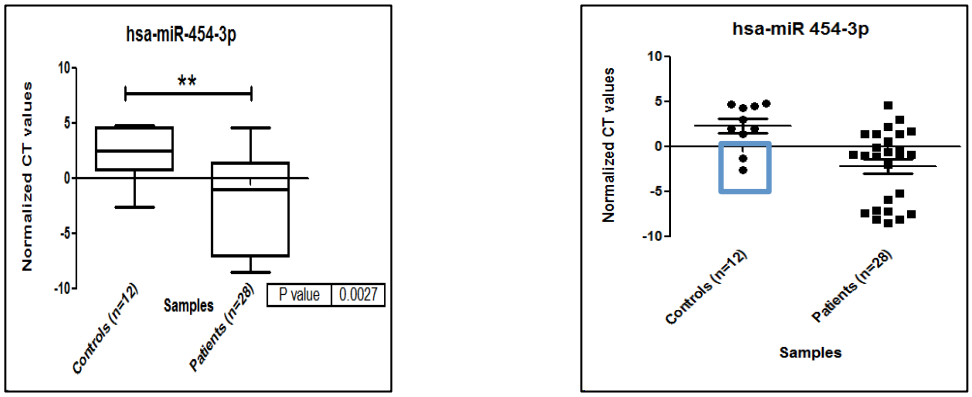


Figure 5-2: Box plots and Dot plots of Normalized CT Values of Dysregulated MiRs: *Box plots and Dot plots of Dysregulated MiRs:* each of these plots displays the p-value associated with the miR.

It is easy to spot two of the outliers in the control group in the dot plots for hsa-miR-127-3p, hsa-miR-204-5p , hsa-miR-329-3p, and hsa-miR-487b-3p **(Figures 5-1B, 5-1D, 5-1F, 5-1H)**. Interestingly, these two samples were consistently the two younger controls (the 24 year-olds) that were included in this study. It is possible that these oncomiRs are being gradually upregulated with age and thus the two youngest samples display the lowest expression levels. Furthermore, these young control samples also showed the greatest expression of tumor suppressor miRs compared to their older counterparts **(Figures 5-1J, 5-1L, and 5-1N)**.

It is also interesting to note that the oldest control (92 years old) and the family history control sample both exhibited intermediate miR profiles more similar to patient samples than controls for some of these miRs. These two samples regularly displayed lower amounts than the rest of the control samples in the dot plots **(Figures 5-1J, 5-1L, and 5-1N)**. Likewise, this same pattern was observed for two oncomiRs **(Figures 5-1F and 5-1H)**. The 92 year-old control and the family history control displayed higher concentrations of these oncomiRs in similar fashion to the patient samples.

The Receiver Operating Curve (ROC) for each of the dysregulated miRs was then generated. This curve provides a graphical plot to illustrate the performance of two groups as the discrimination threshold is varied. The curve is created by plotting the true positive rate (sensitivity) against the true negative rate (1- specificity) at varying thresholds to determine the curve. The sensitivity identifies the proportion of positives that are correctly identified as positives (i.e. a patient sample identified as prostate cancer positive) whereas

the specificity identifies the proportion of negatives that are correctly identified as such (i.e. a control samples identified as prostate cancer negative).

The ROC curve is extremely important because it shows the tradeoff between sensitivity and specificity. As sensitivity increases, the specificity decreases- the ROC curve demonstrates just how great this tradeoff actually is. The Area Under the Curve (AUC) was then calculated to determine how accurate each miR was at discriminating between patients and controls. The greater the area, the more accurate the miR is at discriminating between patients and controls. The AUC of the ROC curve for PSA, the standard measurement for prostate cancer diagnosis, is 0.678. For the seven selected miRs, all the AUCs of their respective ROC curves were significantly higher than that of PSA **(Figure 5-3A through 5-3G)**.

These promising AUCs lead us to believe that constructing a panel of these 7 miRs could prove to be more diagnostic than PS A alone. By using a panel of relevant miRs, a diagnosis of prostate cancer would be best substantiated by the dysregulation of several miRs not based on just one as with PSA.

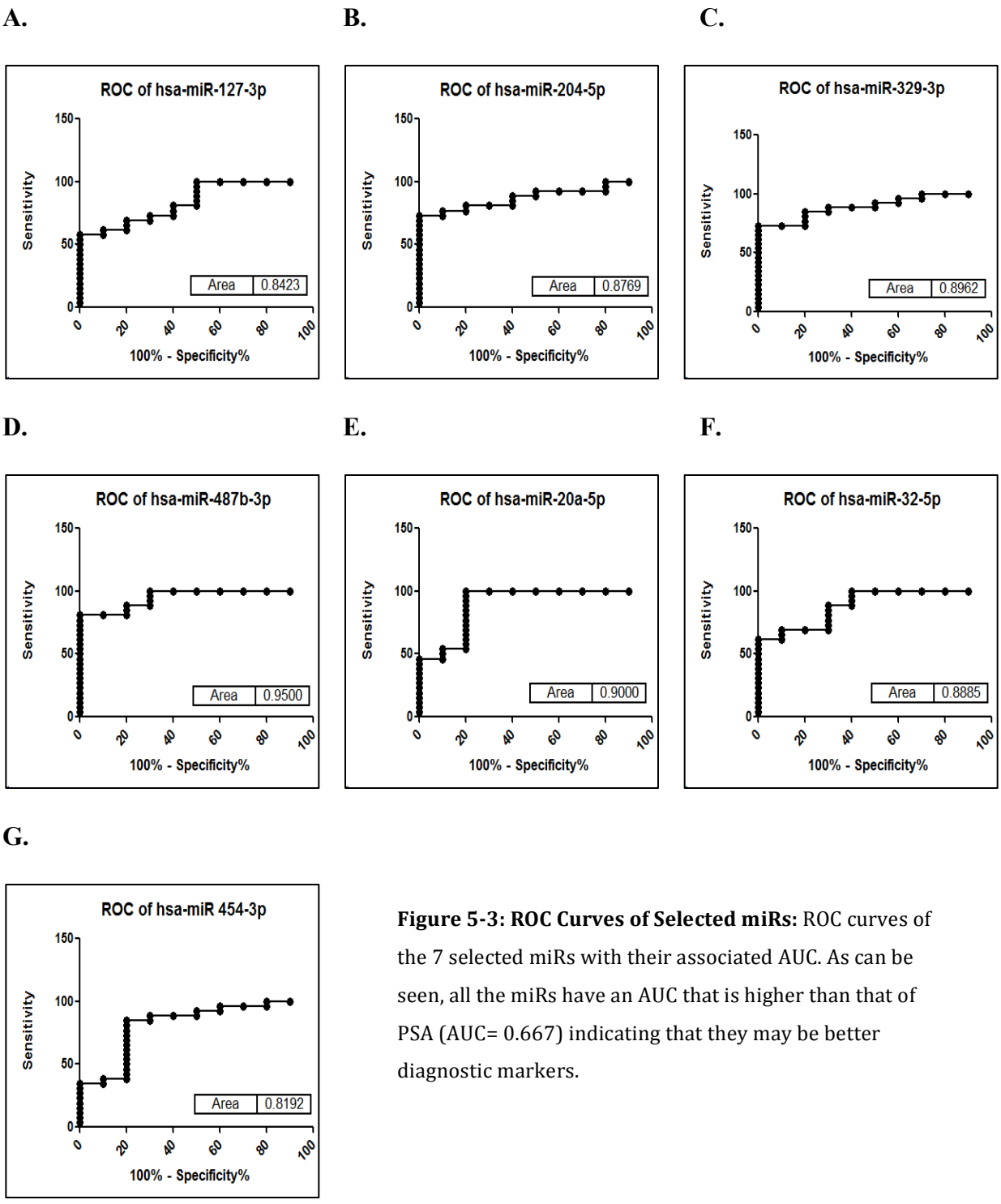


Figure 5-3: ROC Curves of Selected miRs: ROC curves of the 7 selected miRs with their associated AUC. As can be seen, all the miRs have an AUC that is higher than that of PSA (AUC= 0.667) indicating that they may be better diagnostic markers.

Chapter 6

Targets of Dysregulated miRs

Analysis of RNA sequencing results and validation by qRT PCR identified 7 miRs to be of value in diagnosing prostate cancer. All of these miRs generated ROC curves with AUC values considerably better than PSA (AUC=0.667), the current gold standard for diagnosing prostate cancer. A review of these miRs, their chromosomal location, possible targets and a projected role in cancer with particular emphasis on prostate cancer when data is available are discussed below.

I. Upregulated miRs in Blood:

hsa-miR-127-3p:

MiR 127-3p is located on the long arm of chromosome 14. It is situated near a cluster of microRNAs including hsa-mir-431, hsa-mir-433, hsa-mir-432, and hsa-mir-136. This miR is embedded in a CpG island that is susceptible to epigenetic silencing[44].

MiR 127-3p has been found to be lost in glioblastoma proliferation (GBM). GBM is a multistep process during which the expression levels of many genes controlling cell proliferation, cell death, and genetic stability are altered. This study used RNA sequencing on brain tissue and found that miR-127-3p was significantly downregulated in GBM patients. They also found that this microRNA targets SKI, an oncoprotein hence activating TGF- beta[45].

This miR is also downregulated in breast cancer tissue and has been shown to target BCL6. Decreases in miR-127-3p showed marked increases in the oncoprotein BCL6. This protein plays an important role in cell proliferation by suppressing transcription of the anti-apoptotic BCL-XL gene or the adhesion molecule VCAM[46,47]. Therefore, its aberrant expression in cancerous cells promotes tumorigenesis. Chen et al (2013) also showed that over expression of miR-127 or depletion of BCL6 inhibited breast cancer proliferation thereby supporting the tumor suppressor ability of miR-127 on the oncoprotein BCL6[48].

As can be seen in previous results from blood shown above in **Figures 5-1I, 5-2A, 5-2B**, this miR was found to be upregulated in blood from prostate cancer patient samples.

hsa-miR-329-3p:

MiR-329-3p is a nonintronic miR located on the short arm of chromosome 14. It is part of an extensive miR cluster containing over 40 miRs total. One of the miRs in this study, miR-487b-3p, is also part of this extensive cluster and will be discussed more extensively in the next section. Another miR in this cluster, miR-543, located less than 200,000 bases away, was found to be significantly dysregulated with an appreciable p value and FDR value **(Figure 5-1J)** and will be studied at a later time.

In a study conducted by Yang et al (2014), miR-329-3p was found to be downregulated in neuroblastoma metastatic tumor tissue compared to the primary tumor. When SH-SY5Y cells, a human neuroblastoma cell line, were transfected with miR-329, the growth and motility of the cells was significantly suppressed and prevented colony formation. They also conducted studies to determine potential targets of this miR. One promising target was KDM1A, which has been shown to be a tumorigenic protein[49].

Another research group studying recurrent prostate cancer found that KDM1A was significantly upregulated in the LnCaP androgen dependent prostatic cell lines. When this protein was depleted using siRNA, VEGF-A expression was also decreased which in turn blocked androgen induced VEGF-A, PSA, and Tmprss2 expression[50].

Interestingly, a study on metastatic prostate cancer cells found that 10 miRs in this cluster, including miR 329 and miR 487b were significantly downregulated. This study also found an inverse correlation between increased Gleason scores and metastases versus decreases in expression. In addition, these miRs were found to play an important role in

regulating proliferation, apoptosis, migration and invasion in metastatic prostate cancer cells.

Once again, the inverse is seen in blood from prostate cancer patients with a significant upregulation in miR 329-3p as opposed to what is detected in various tissue and prostate cell lines.

hsa-miR-487b-3p:

MiR 487 is located almost 15 kb way from miR-329 on the same microRNA cluster on chromosome 14. This miR, like miR-329, has been found to be downregulated in prostate cancer cell lines[51]**.** Another study by Gattolliat et al, also showed that miR-487b-3p was downregulated by 5.75 fold in neuroblastomas tissue samples[52].

An interesting predicted target for miR-487b-3p was ALDH1A3, aldehyde dehydrogenase 1A3, an enzyme known to be upregulated in the LnCaP prostate cancer cell lines. When LnCaP cells are exposed to the androgen dihydrotestosterone (DHT), this protein increased 4-fold. This protein has also been implicated in breast cancer metastases and is therefore notably an oncomiRic protein[53].

Once again, the literature shows an inverse relationship to what the RNA sequencing and confirmatory qRT-PCR data on whole blood samples has shown in this study.

hsa-miR-204-5p:

MiR-204-5p is located on chromosome 9 in the sixth intron of its host gene TRPM3, transient receptor potential cation channel subfamily M member 3. TRPM3 is important for cellular calcium signaling and maintaining homeostasis. Interestingly, miR-204-5p is located in a cancer- associated genomic region in the long arm of chromosome 9 and has been shown to be extremely downregulated in many tumor types including breast, kidney and prostate[54].

Wang et al (2010) had previously shown that reduction in the expression of miR-204- 5p led to reduce expression of claudins 10,16, and 19, which decreased transepithelial resistance by 80% and reduced cell membrane voltage and conductance. They also showed that absence of 204-5p also led to decrease in Kir7.1 proteins, which connect TGF-BR2 and maintain potassium homeostasis. Therefore this miR plays a crucial role in maintaining epithelial barrier function and cell physiology[54]. Another study showed that 204-5p suppressed the growth, migration and invasion of endometrial carcinomas by binding TrkB mRNA and interfering with JAK2 and STAT3 phosphorylation[55].

Similarly to miR-127-3p, miR-204-5p was also upregulated in prostate cancer patient blood samples, opposite to what has been shown in the literature when dealing with tissue samples.

II. Downregulated miRs in Blood:

hsa-miR-20a-5p:

MiR-20a-5p, located on chromosome 13, is part of the 17/92 cluster, which has been implicated in age related diseases. It has been found that this cluster is vital to normal development and plays an important role in cell cycle, proliferation, apoptosis, and other processes. One of the most studied targets of miR-20a is the E2F family, particularly E2F2 and E2F3, which play a role in cell cycle regulation. This miR also targets several cyclin dependent kinases, including p21 and p57, which halt the cell cycle in case of aberrant division. Finally, another notable target is FasI, which promotes cell death. Interestingly, the major targets of this miR promote tumorigenesis and angiogenesis by preventing cell cycle checkpoints[56].

One study using tumor tissues from prostate cancer patients found that MiR-20a was significantly higher in dedifferentiated CaP cells. The over expression of this MiR has also been shown in the PC3 prostate cancer cell lines using PCR and microarray studies on tissue[57]. Once again, this study found that this miR binds E2F2 and E2F3 mRNAs in the 3' UTR and play a crucial role in the cell cycle[58].

As with the rest of the of miRs, our results from the RNA sequencing and confirmatory qRT-PCR data tell the opposite story for this miR. It has been found to be downregulated in blood samples from patients.

hsa-miR-32-5p:

Finally, miR-32-5p, located on chromosome 9, has been found to be an androgen-regulated miR that targets BTG2. This miR is overexpressed in castration- resistant prostate cancer (CRPC). An increase in 32-5p expression levels was associated with reduced apoptosis in tumor samples by targeting the 3' UTR of BTG2 and preventing apoptosis[59].

Furthermore, this miR was discovered to be regulated by DHT and showed that putative Androgen Receptor- binding sites (ARBSs) upstream without other genes nearby.

This miR was also found to be downregulated in patient blood samples.

hsa-miR-454-3p:

MiR- 454-3p is located on chromosome 17 in the first intronic region of SKA2, its host gene. SKA2, also known as Spindle and Kinetochore Associated Complex Subunit 2, is essential for proper chromosome segregation. During the cell cycle both, SKA2 and miR 454-3p have been shown to be upregulated.

MiR 454-3p has been shown to target the B cell Translocation Gene 1 (BTG1) that has long been recognized as a tumor suppressor gene. BTG1 plays an important role in cell cycle progression and is involved in stress responses. This anti-proliferative gene is expressed at its highest concentration during the G0/G1 phases of the cell cycle and then downregulated when the cell progresses through G1 phase[60].

A study showed that in renal carcinoma cells, an increase in 454-3p displayed a marked decrease in BTG1 via a direct interaction with the 3' UTR of BTG1 mRNA. Furthermore, their cell cycle analysis showed that when 454-3p was overexpressed, it shifted the cell cycle arrest from G2/M into S phase[60].

Interestingly, a second study also found miR-454-3p was overexpressed in chronic myeloid leukemia by targeting TGFBR2 (Transforming Growth Factor, beta Receptor II), a member of the Ser/Thr protein kinase family and the TGFB receptor subfamily. This receptor/ligand complex phosphorylates proteins that enter the nucleus and regulates the transcription of genes related to cell proliferation. An increase of miR-454-3p in myeloid leukemia led to a decrease in this receptor[61].

As with the rest of the miRs, this miR showed the inverse expression in blood as to what is shown in the literature. RNA sequencing and confirmatory qRT-PCR data showed that this miR was downregulated in patient blood samples.

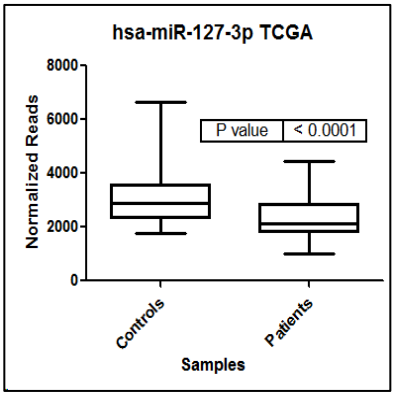
III. Comparison of miR expression in Blood to the TGCA Database:

Remarkably, the literature showed an opposite result to what our RNA sequencing data and confirmatory qRT-PCR displayed. In an effort to better characterize the role of these 7 miRs in prostate cancer, a comparison was made to data from the Cancer Genome Atlas (TCGA) database. Here, RNA sequencing data preformed on 495 prostate cancer tumor samples as well as tumor-free margins from the same patient was deposited into a database as overseen by the National Cancer Institute (NIH) and made freely available for review by investigators.

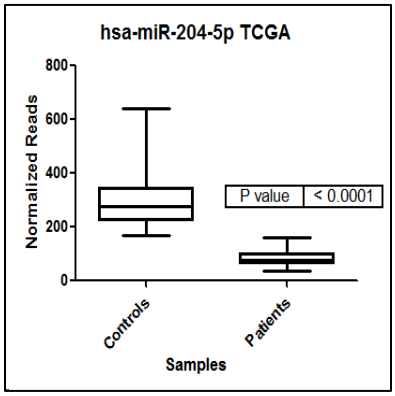
Interestingly, this comparison also showed an inverse relationship between miR expression levels detected in blood compared to actual prostate tumor tissue. For example, miRs that were found to be upregulated in our blood samples were found to be tumor suppressors lost in tumors in the TCGA data. The opposite was also true, miRs downregulated in blood were found to be aberrantly overexpressed as oncomiRs in tumor tissue.

A search of the literature found that several studies have found a similar comparison to be true for many miRs, although not all. This raises the question of whether oncomiRs in tumor samples were being retained within the tumor for the benefit of tumor growth and therefore what escapes into the blood is decreased. On the other hand, tumor suppressors that could block tumor growth were being dumped out into the blood.

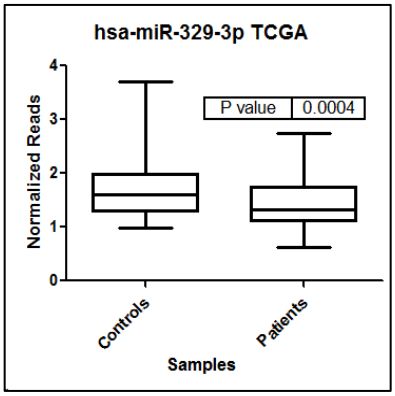
A.


hsa-miR-127-3p TCGA
Normalized Reads
8000
6000
4000
2000
0
P value
< 0.0001
Controls
Patients
Samples

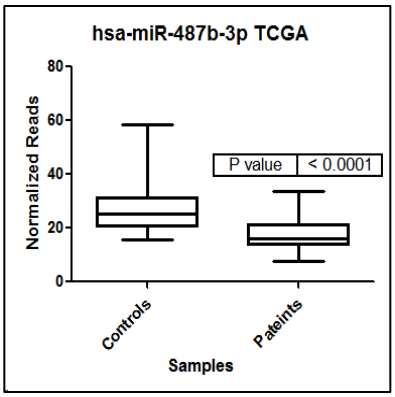
B.


hsa-miR-204-5p TCGA
Normalized Reads
800
600
400
200
0
P value
< 0.0001
Controls
Patients
Samples

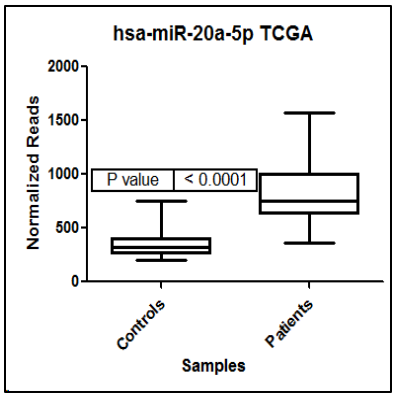
C.


hsa-miR-329-3p TCGA
Normalized Reads
4
3
2
1
0
P value
0.0004
Controls
Patients
Samples

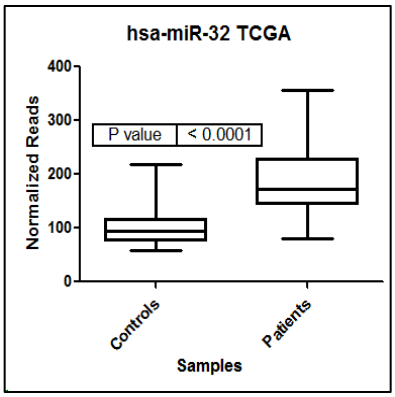
D.


hsa-miR-487b-3p TCGA
Normalized Reads
80
60
40
20
0
P value
< 0.0001
Controls
Pateints
Samples

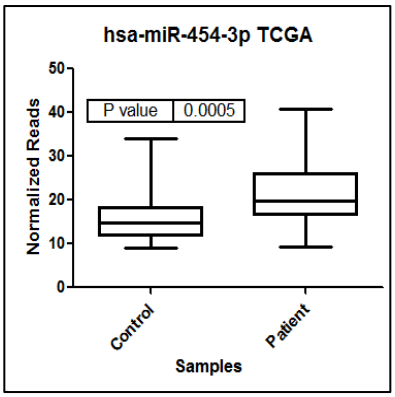
E.


hsa-miR-20a-5p TCGA
Normalized Reads
2000
1500
1000
500
0
P value
< 0.0001
Controls
Patients
Samples

F.


hsa-miR-32 TCGA
Normalized Reads
400
300
200
100
0
P value
< 0.0001
Controls
Patients
Samples

G.


hsa-miR-454-3p TCGA
Normalized Reads
50
40
30
20
10
0
P value
0.0005
Control
Patient
Samples

Chapter 7

Conclusions

RNA sequencing of microRNAs is a fairly new technique and none of the previous studies in the literature analyzed individual miR profiles of patient and control samples. All these studies pooled samples together in order to obtain sufficient material for RNA sequencing. This method prevents an analysis of variability across individual sample profiles and blocks any correlation to the stage of disease when Gleason scores are available. Not only is it important to diagnose prostate cancer, but also identify biomarkers that can serve to stage disease or separate indolent from aggressive disease which is even more important to subsequent treatment options. Therefore, the fact that our RNA sequencing data examined 40 samples individually without pooling, enabled us to analyze trends those other studies were not able to evaluate. It also permitted us to determine outliers and hypothesize why those outliers might act differently than the rest of the group.

Furthermore, the lack of reliable normalization candidates has been crippling to many studies. The RNA sequencing data allowed us to select miRs that were consistent across all 40 samples and report reliable normalization miRs that can be used for the analysis of qRT-PCR in blood. The geometric mean of miR-197-3p, miR-1538, miR-1468-3p, and miR-146a-5p has shown to be remarkably consistent among all the patients and controls, thus providing reliable miRs for normalization.

It was important to confirm RNA sequencing results with qRT-PCR. Significantly, results from these two very different methodologies agreed well further supporting the quality of our approach. All the miRs with low p-values and low FDR

values in the RNA sequencing data showed extraordinary p-values with qRT-PCR and astonishing AUCs. This was extremely encouraging because thus far investigators have been determining which miRs to study by reviewing the literature. Thus, they are limited to analyzing only those miRs that have already been shown to be dysregulated in disease. However, RNA sequencing allows for the identification of all possible diagnostic miRNAs both known and novel, expanding the spectrum of miRs that can contribute to disease.

Here, an interesting discovery was the inverse relationship between blood miR expression levels and tumor expression levels. Since all of our blood miRs displayed an inverse expression level with tumor samples, we propose that it is possible that tumors are retaining oncomiRs for the purpose of driving tumorigenesis and angiogenesis and therefore, less of these oncomiRs are released into the blood. Conversely, tumor suppressors block tumor growth and may need to be disposed of to enhance tumorigenesis and ultimately metastasis, hence the increase in blood levels of these miRs. If this was the case for only one or two miRs, perhaps not, but recurrence supports this hypothesis.

It has been shown that cancer cells secrete vesicles containing not only mature miRs to modify their environment for future metastasis, but the entire processing machinery (dicer, RISC with premiR) to ensure that once taken up by the target cell, the miR is efficiently processed and actively moved into translational silencing of target mRNAs. It is felt that if only the mature miR was delivered, it might not be as efficient in modifying translation within the target cell. With this in mind, it is not much of a stretch to propose a developing tumor wants to dispose of compromising miRs that could restrict

its growth; thus, excluding tumor suppressor miRs. Concomitantly, holding onto an oncomiR to quickly modulate the proteome is fast and efficient, faster than modifying gene expression at the transcriptional level. Similar studies to ours for other cancers need to be completed to determine the overall merits of this hypothesis.

Since three of the miRs on our panel belong to the same mega cluster on chromosome 14, it would also be interesting to evaluate more of the miRs from this cluster. After a more recent review of our data, we found several additional miRs from this cluster with low p-values and FDR values, which need to be further evaluated. It would be interesting to see if these miRs are also increased in patient blood samples as opposed to the TCGA tumor samples further supporting our hypothesis. Studies have shown that the miRs in this cluster are downregulated during tumorigenesis through mechanisms that are unknown. If increases in these miRs are found in blood, it is possible to hypothesize that the expression of these miRs is not being turned off at the transcriptional level but that they are being shuttled out of the tumor cell and into the blood as a survival and growth mechanism.

In summary, the 7 miR panel that we have developed shows that miR-127-3p, miR-329-3p, miR-487b-3p, miR-204-5p, miR-20a-5p, miR-454-3p, and miR-32-5p display higher diagnostic capabilities than PSA measurements. ROC curves and AUC values are outstanding compared to the current PSA gold standard. Further studies need to be conducted to determine a threshold level for each of these miRs that can diagnose prostate cancer.

The four upregulated miRs in patient blood (miR-127-3p, miR-329-3p, miR-487b-3p, and miR-204-5p) that cumulatively target BCL6, KDM1A, ALDH1A3, and TrkB which have been shown to be important regulators of prostate cancer. These proteins have been shown to have oncogenic effects on tissue and therefore these miRs are tumor suppressors that are lost in tumorigenesis. On the other hand, the three downregulated miRs in patient blood (miR-20a-5p, miR-32-5p, and miR-454-3p) have been shown to target the tumor suppressor proteins E2F2, BTG2, and BTG1 respectively. Although they were downregulated in our patient blood samples, they have been shown to be oncomiRs in tumor tissue.

Another value of this suggested panel is that the four upregulated miRs in blood and 3 downregulated miRs in blood serve as internal controls for each other thereby serving to verify the accuracy of the panel results (i.e. they don't all go up or all go down). Constructing a panel with only downregulated miRs is always hard to justify however by pairing the loss of these three miRs with the increase in the other four miRs allows for greater diagnostic confidence.

References

1. Wilson AH. The prostate gland: a review of its anatomy, pathology, and treatment. *JAMA*. 2014 Aug 6; 312(5): 562

2. Huggins C, Scott WW, Heinen JH. Chemical composition of human semen and of the secretions of the prostate and seminal vesicles. *American Journal of Physiology*. 1942; 136(3): 467-473.

3. Kavanagh JP. Isocitric and citric acid in human prostatic and seminal fluid: implications for prostatic metabolism and secretion. *Prostate*. 1994; 24(3): 139-42.

4. Kumar VL, Majumder PK. Prostate Gland: Structure, Functions and Regulation. *International Urology and Nephrology*. 1995; 27(3): 231-243.

5. Oh WK, Hurwitz M, D'Amico AV, et al. Biology of Prostate Cancer. In: Kufe DW, Pollock RE, Weichselbaum RR, et al., editors. Holland-Frei Cancer Medicine. 6th edition. Hamilton (ON): BC Decker; 2003.

6. National Cancer Institute. SEER Training Modules. Lobes of the Prostate. http://training.seer.cancer.gov/prostate/anatomy/lobes.html. Accessed January 15, 2015.

7. Lija H, Ulmert D, Vickers AJ. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer*. 2008 Apr;8(4): 268-78.

8. Balk SP, Ko YJ, Bubley GJ. Biology of Prostate- Specific Antigen. *J Clin Oncol*. 2003 Jan 15; 21(2): 383-91.

9. Diamandis EP, Yousef GM. Human tissue kallikreins: a family of new cancer biomarkers. *Clin* Chem. 2002 Aug; 48(8):1198-205.

10. Witte JS, Goddard KAB, Conti DV, et al. Genomewide scan for prostate cancer-aggressiveness loci. *Am J Hum Genet*. 2000 Jul; 67(1): 92-99.

11. Slager SL, Schaid DJ, Cunningham JM. Confirmation of linkage of prostate cancer aggressiveness with chromosome 19q. *Am J Hum Genet*. 2003 Mar; 72(3): 759-762

12. National Cancer Institute. Prostate- Specific Antigen (PSA) http://www.cancer.gov/types/prostate/psa-fact-sheet. Accessed December 13, 2014.

13. Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate- specific antigen level < or =4.0 ng per milliliter. *N Engl J Med*. 2004 May 27;350(22):2239-46.

14. Luboldt H, Schindler JF, Rübben H. Age-specific ranges for PSA in the detection of prostate cancer. *Oncology*. 1997 April; 11(4): 475-82, 485; discussion 485-6, 489.

15. American Cancer Society. Diagnosis of Prostate Cancer. http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-diagnosis. Accessed December 13, 2014.

16. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. *Mod Pathol*. 2004 Mar; 17(3): 292-306.

17. Ha M, Kim VN. Regulation of microRNA biogenesis. *Nature Reviews Mol Cell Bio.* 2014 Jul 16; 15, 509-524.

18. Appasani Krishnarao. MicroRNAs in disease biology. Book

19. Baek D, Villén J, Shin C, et al. The impart of microRNAs on protein output. *Nature*. 2008 September 4; 455, 64-71

20. Wang X. Composition of seed sequence is major determinant of microRNA targeting patterns. *Bioinformatics*. 2014 May 15; 30(10): 1377-83

21. Voss T, Wengeler M, Heese F, Wyrich R. Isolation of small RNA species from PAXgene Blood RNA Tubes. *Qiagen*.

22. Chen X, Ba Y, Ma L, et al. Characterization of microRNAs in serum: A novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res.* 2008;18(10):997-1006.

23. Etheridge A, Lee I, Hood L, Galas D, Wang K. Extracellular microRNA: A new source of biomarkers. *Mutat Res.* 2011;717(1-2):85-90.

24. Lasser C, Alikhani VS, Ekstrom K, et al. Human saliva, plasma and breast milk exosomes contain RNA: Uptake by macrophages. *J Transl Med*. 2011;9:9-5876-9-9.

25. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, Lotvall JO. Exosome- mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol*. 2007;9(6):654-659.

26. Katahira J, Yoneda Y. Nucleocytoplasmic transport of microRNAs and related small RNAs. *Traffic*. 2011;12(11):1468-1474.

27. Creemers EE, Tijsen AJ, Pinto YM. Circulating MicroRNAs. *Circulation Research.* 2012; 110: 483-495.

28. Schultz NA, Dehlendorff C, Jensen BV, et al. MicroRNA Biomarkers in Whole Blood for Detection of Pancreatic Cancer. *JAMA.* 2014;311(4):392-404.

29. Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for prostate cancer detection. *Proc Natl Acad Sci U S A*. 2008 Jul 29; 105(30): 10513–10518.

30. Pritchard CC, Cheng HH, Twari Muneesh. MicroRNA profiling: Approaches and Conciderations. *Nature Reviews Genetics* 13, 358-369.

31. Ardekani AM, Naeini MM. The role of microRNAs in human disease. *Avicenna J Med Biotechnol*. 2010 Oct-Dec; 2(4): 161-179.

32. Budd WT. Combinatorial analysis of tumorigenic microRNAs driving prostate cancer. [dissertation]. Virginia Commonwealth University; 2012.

33. Seashols SJ. Variation and Modulation of microRNAs in Prostate Cancer and Biological Fluids. [dissertation]. Virginia Commonwealth University; 2013.

34. Fletcher CE, Dart DA, Sita- Lumsden A. Androgen- regulated processing of the oncomir miR- 27a, which targets Prohibitin in prostate cancer. *Hum Mol Genet*. 2012 Jul 15; 21(14): 2112-27

35. Munch EM, Harris RA, Mohammad M, et al. Transcriptome Profiling of microRNA by Next-Gen Deep Sequencing Reveals Known and Novel miRNA Species in the Lipid Fraction of Human Breast Milk. *Plos One.* 2013 Feb 13; 505(8): 10513–10518.

36. Robinson, MD, and Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2012:11; R25

37. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008, 5:621-628

38. Gene Specfic Analysis White Paper. Partek.com. Date of access: May 3, 2015

39. Gee HE, Buffa FM, Camps C, et al. The small-nucleolar RNAs commonly used for microRNA normalisation correlate with tumour pathology and prognosis. *Br J Cancer*. 2011 Mar 29;104(7):1168-77.

40. Schaefer A, Jung M, Miller K, et al. Suitable reference genes for relative quantification of miRNA expression in prostate cancer. *Exp Mol Med*. 2010 Nov 30;42(11):749-58.

41. Egidi MG, Cochetti G, Guelfi G, et al. Stability Assessment of Candidate Reference Genes in Urine Sediment of Prostate Cancer Patients for miRNA Applications. *Disease Markers*. 2015 April; 2015;2015

42. Mestdagh P, Vlierberghe PV, Weer AD, et al. A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biology* 2009, 10:R64.

43. Peltier HJ, Latham GJ. Normalization of microRNA expression levels in quantitative RT-PCR assays: Identification of suitable reference RNA targets in normal and cancerous human solid tissues. *RNA*. 2008 May;14(5):844-52.

44. Lopez-Serra P, Esteller M. DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer. *Oncogene.* 2012 Mar 29;31(13):1609-22.

45. Huawei Jiang, Chengmeng Jin, Jie Liu, et al. Next Generation Sequencing Analysis of miRNAs: MiR-127-3p Inhibits Glioblastoma Proliferation and Activates TGF-β Signaling by Targeting SKI. *OMICS*. 2014 Mar 1; 18(3): 196–206.

46. Tang TT, Dowbenko D, Jackson A, et al. The Forkhead transcription factor AFX activates apoptosis by induction of the Bcl-6 transcriptional repressor. *J Biol Chem.* 2002;277:14255-65.

47. Mencarelli A, Renga B, Distrutti E, Fiorucci S. Antiatherosclerotic effect of farnesoid X receptor. *Am J Physiol Heart Circ Physiol.* 2009;296:H272-81.

48. Jingwen Chen, Miao Wang, Mingzhou Guo, et al. miR-127 Regulates Cell Proliferation and Senescence by Targeting BCL6. *PLoS One*. 2013; 8(11): e80266.

49. Yang H, Li Q, Zhao W, et al. miR-329 suppresses the growth and motility of neuroblastoma by targeting KDM1A. *FEBS Lett.* 2014 Jan 3;588(1):192-7.

50. Kashyap V, Ahmad S, Nilsson EM, et al. The lysine specific demethylase-1 (LSD1/KDM1A) regulates VEGF-A expression in prostate cancer. *Mol Oncol*. 2013 Jun;7(3):555-66.

51. Formosa A, Markert EK, Lena AM, et al. MicroRNAs, miR-154, miR-299-5p, miR-376a, miR-376c, miR-377, miR-381, miR-487b, miR-485-3p, miR-495 and miR-654-3p, mapped to the 14q32.31 locus, regulate proliferation, apoptosis, migration and invasion in metastatic prostate cancer cells. *Oncogene.* 2014 Oct 30;33(44):5173-82.

52. Gattolliat CH, Thomas L, Ciafrè SA, et al. Expression of miR-487b and miR-410 encoded by 14q32.31 locus is a prognostic marker in neuroblastoma. *Br J Cancer.* 2011 Oct 25;105(9):1352-61.

53. Le Magnen C, Bubendorf L, Rentsch CA, et al. Characterization and clinical relevance of ALDHbright populations in prostrate cancer. *Clin Cancer Res*. 2013 Oct 1;19(19):5361-71.

54. Wang FE, Zhang C, Maminishkis A, et al. MicroRNA-204/211 alters epithelial physiology. *FASEB J*. 2010 May;24(5):1552-71.

55. Bao W, Wang HH, Tian FJ, et al. A TrkB-STAT3-miR-204-5p regulatory circuitry controls proliferation and invasion of endometrial carcinoma cells. *Mol Cancer.* 2013 Dec 9;12:155.

56. E Mogilyansky and I Rigoutsos. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death and Differentiation* (2013) 20, 1603–1614.

57. Sylvestre Y, De Guire V, Querido E, et al. An E2F/miR-20a autoregulatory feedback loop. *J Biol Chem*. 2007 Jan 26;282(4):2135-43.

58. Pesta M1, Klecka J, Kulda V, et al. Importance of miR-20a expression in prostate cancer tissue. *Anticancer Res.* 2010 Sep;30(9):3579-83.

59. Jalava SE, Urbanucci A, Latonen L, et al. Androgen-regulated miR-32 targets BTG2 and is overexpressed in castration-resistant prostate cancer. *Oncogene.* 2012 Oct 11;31(41):4460-71.

60. Wu X, Ding N, Hu W, et al. Down-regulation of BTG1 by miR-454-3p enhances cellular radiosensitivity in renal carcinoma cells. *Radiat Oncol*. 2014 Aug 12;9:179.

61. Xiong Q, Yang Y, Wang H, et al. Characterization of miRNomes in acute and chronic myeloid leukemia cell lines. *Genomics Proteomics Bioinformatics.* 2014 Apr;12(2):79-91.